HANDBOOKS IN ECONOMICS 2

HANDBOOK OF ECONOMETRICS

VOLUME 6A

Editors:
James J. Heckman
Edward E. Leamer

NORTH-HOLLAND

CONTENTS OF THE HANDBOOK

VOLUME 1

Part 1: MATHEMATICAL AND STATISTICAL METHODS IN ECONOMETRICS

Part 2: ECONOMETRIC MODELS

Part 3: ESTIMATION AND COMPUTATION

VOLUME 4

Part 9: ECONOMETRIC THEORY

PREFACE TO THE HANDBOOK

As conceived by the founders of the Econometric Society, econometrics is a field that uses economic theory and statistical methods to address empirical problems in economics. It is a tool for empirical discovery and policy analysis. The chapters in this volume embody this vision and either implement it directly or provide the tools for doing so. This vision is not shared by those who view econometrics as a branch of statistics rather than as a distinct field of knowledge that designs methods of inference from data based on models of human choice behavior and social interactions. All of the essays in this volume offer guidance to the practitioner on how to apply the methods they discuss to interpret economic data. The authors of the chapters are all leading scholars in the fields they survey and extend.

Auction theory and empirical finance are two of the most exciting areas of empirical economics where theory and data combine to produce important practical knowledge. These fields are well represented in this Handbook by Susan Athey and Philip Haile (auctions) and Lars Hansen, John Heaton, Nikolai Roussanov and Junghoon Lee (finance). Both papers present state of the art knowledge of their respective fields and discuss economic models for the pricing of goods and risk. These papers feature agent response to uncertainty as an integral part of the analysis. Work on the pricing of labor services lies at the core of empirical labor economics. Thomas MaCurdy surveys empirical methods for estimating wage equations from panel data in a way that is accessible to practitioners.

The econometrics of industrial organization (IO) is another vibrant area of applied econometrics. Scholars in the field of IO have embraced econometrics. The resulting symbiosis between theory and practice is a paragon for econometric research. Modern developments in game theory have been incorporated in econometric models that enrich both theory and empirical analysis. These developments are well-represented in this volume by the essays of Daniel Ackerberg, Lanier Benkard, Steven Berry, and Ariel Pakes and of Peter Reiss and Frank Wolak. Stephen Bond and John van Reenen summarize the related literature on modeling the dynamics of investment and employment, which is an integral part of macroeconomics and modern IO.

The essay by Erwin Diewert and Alice Nakamura surveys methods for measuring national productivity. They exposit a literature that provides the tools for comparing the economic performance of policies and of nations. The authors survey the methods that underlie this important field of economics. Edward Leamer's essay stresses the interplay between data and theory in the analysis of international trade patterns. In an increasingly global market, the measurement of trade flows and the study of the impact of trade on economic welfare is important for understanding recent economic trends.

Modern economics has come to recognize heterogeneity and diversity among economic agents. It is now widely acknowledged that the representative agent paradigm is an inaccurate and misleading description of modern economies. The essay by Richard Blundell and Thomas Stoker summarizes and synthesizes a large body of work on the aggregation of measurements across agents to produce reliable aggregate statistics and the pitfalls in the use of aggregates.

Consumer theory, including the theory of labor supply, is at the heart of empirical economics. The essay by Richard Blundell, Thomas MaCurdy, and Costas Meghir surveys a vast literature with an ancient lineage that has been at the core of empirical economics for over 100 years. They develop empirical models of consumer demand and labor supply in an integrated framework.

The evaluation of economic and social programs is a central activity in economics. It is the topic of three essays in this Handbook. James Heckman and Edward Vytlacil contribute two chapters. The first chapter moves the literature on program evaluation outside of the framework of conventional statistics to consider economic policy questions of interest, to incorporate agent choice behavior and the consequences of uncertainty, and to relate the recent work on policy evaluation in statistics to older and deeper frameworks developed in econometrics. Issues of causality and the construction of counterfactuals are addressed within the choice-theoretic framework of economics.

Their second chapter uses the *marginal treatment effect* to unify a diverse and disjointed literature on treatment effects and estimators of treatment effects. The marginal treatment effect can be interpreted as a willingness to pay parameter. This chapter focuses on mean treatment effects in static environments without explicit analysis of uncertainty.

The essay by Jaap Abbring and James Heckman surveys new methods for identifying *distributions* of treatment effects under uncertainty. It surveys and develops methods for the analysis of dynamic treatment effects, linking the statistical literature on dynamic sequential randomization to the econometric literature on dynamic discrete choices. It also surveys recent approaches to the general equilibrium evaluation of social programs.

One of the most important contributions of econometric theory to empirical knowledge is the analysis of the identifiability of econometric models – determining under what conditions a unique model describes the data being used in an empirical analysis. Cowles Commission analysts formalized these ideas, focusing largely on linear systems [Tjalling Koopmans, Herman Rubin, and Roy Leipnik (1950)]. Later work by Franklin Fisher (1966) extended the Cowles analysis to nonlinear, but parametric systems. Rosa Matzkin's contribution to this Handbook synthesizes and substantially extends these analyses to consider a large body of work on the identification of nonparametric models. The methods she surveys and extends underlie a large literature in applied economics.

Hidehiko Ichimura and Petra Todd present a guide to the recent literature on nonparametric and semiparametric estimators in econometrics that has been developed in

the past 20 years. They conduct the reader through the labyrinth of modern nonparametric econometrics to offer both practical and theoretical guides to this literature.

Robert Moffitt and Geert Ridder address the important problem of how to combine diverse data sets to identify models and improve the precision of estimation of any model. This topic is of great importance because many data sets in many areas of economics contain valuable information on subsets of variables which, if they were combined in a single data set, would identify important empirical relationships. Moffitt and Ridder present the state of the art in combining data to address interesting economic questions.

Xiaohong Chen presents a detailed, informative survey of sieve estimation of semiparametric models. The sieve principle organizes many different approaches to nonparametric and semiparametric estimation within a common analytical framework. Her analysis clarifies an extensive and widely used literature. Marine Carrasco, Jean-Pierre Florens, and Eric Renault survey the literature on nonparametric and semiparametric econometrics that is based on inverse operators. Their analysis subsumes recent research on nonparametric instrumental variable methods as well as research on deconvolution of distributions. They present both theoretical and practical guides to this frontier area of econometrics.

JAMES J. HECKMAN
University of Chicago, Chicago, USA
American Bar Foundation, USA
University College Dublin, Dublin, Ireland

EDWARD E. LEAMER
University of California, Los Angeles, USA

Acknowledgements

We gratefully acknowledge support from the National Science Foundation, the University of Chicago, and University College London for conferences at which many of these papers were presented. We also thank the many referees and conference participants whose helpful comments have improved every chapter in this volume.

References

Fisher, F.M. (1966). The Identification Problem in Econometrics. McGraw–Hill, New York.

Koopmans, T.C., Rubin, H., Leipnik, R.B. (1950). "Measuring the equation systems of dynamic economics". In: Koopmans, T.C. (Ed.), Statistical Inference in Dynamic Economic Models. In: Cowles Commission Monograph, Number 10. John Wiley & Sons, New York, pp. 53–237. Chapter 2.

Chapter 60

NONPARAMETRIC APPROACHES TO AUCTIONS *

SUSAN ATHEY
Harvard University, USA
NBER, USA
e-mail: athey@fas.harvard.edu

PHILIP A. HAILE
Yale University, USA
NBER, USA
e-mail: philip.haile@yale.edu

Contents

* We have benefited from the input of Jaap Abbring, Estelle Cantillon, Ali Hortaçsu, Elena Krasnokutskaya, Jonathan Levin, Harry Paarsch, Isabelle Perrigne, Martin Pesendorfer, Joris Pinkse, Rob Porter, Unjy Song, Elie Tamer, an anonymous referee, and students at Stanford, University College London, and Yale. Ying Fan, Dan Quint and Gustavo Soares provided capable research assistance. Research support from the National Science Foundation (grants SES-0112047 and SES-0351500) and the Alfred P. Sloan Foundation is gratefully acknowledged. Any conclusions, findings, or opinions are those of the authors and do not necessarily reflect the views of any funding organization.

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06060-6

Abstract

This chapter discusses structural econometric approaches to auctions. Remarkably, much of what can be learned from auction data can be learned without restrictions beyond those derived from the relevant economic model. This enables us to take a nonparametric perspective in discussing how the structure of auction models can be combined with observables to uncover (or test hypotheses about) primitives of interest in auction markets. We focus on first-price sealed-bid and ascending auctions, including extensions to Dutch auctions, Internet auctions, multi-unit auctions, and multi-object auctions. We consider a wide range of underlying structures of bidder demand and information, as well as a variety of types of data one may encounter in applications. We discuss identification and testable restrictions of these models and present a variety of estimation approaches.

Keywords

auctions, identification, estimation, testing

JEL classification: C5, C14, D44

1. Introduction

Auctions provide opportunities for economists to examine field data from markets that can involve rich strategic interaction and asymmetric information while nonetheless being simple enough that the salient forces can be convincingly captured by a tractable economic model. The primitives of any strategic model include the set of players, the information structure, the rules of play, and players' objectives. In auction markets, one can often describe these key elements with an unusually high degree of confidence. Consequently, auctions have been at the center of efforts to combine economic theory with econometric analysis to understand behavior and inform policy.

Early work by Hendricks and Porter (1988) and others played an important role in demonstrating the empirical relevance of private information and the ability of strategic models to predict behavior. More recently, there has been a great deal of attention to econometric approaches to auctions that incorporate restrictions from economic theory as assumptions of an econometric model.[1] The goal of this structural approach is to address questions that can only be answered with knowledge concerning the distribution functions that characterize the underlying demand and information structure. Structural empirical work on auctions has examined, for example, the division of rents in auctions of public resources, whether reserve prices in government auctions are adequate, the effects of mergers on procurement costs, whether changes in auction rules would produce greater revenues, whether bundling of procurement contracts is efficient, the value of seller reputations, the effect of information acquisition costs on bidder participation and profits, whether bidders' private information introduces adverse selection, and whether firms act as if they are risk averse.

Many of these questions have important implications well beyond the scope of auctions themselves. In all of economics there is a tradeoff between the assumptions one relies on and the questions one can address. Because an auction is a market institution that is particularly easy to capture with a theoretical model, one may have more confidence than usual that imposing significant structure from economic theory in interpreting data can be useful. Combined with the fact that private information and strategic behavior are paramount in auctions, this suggests that auctions may enable economists to get at questions of importance to many other types of markets.

Remarkably, much of what can be learned from auction data can be learned without restrictions beyond those derived from economic theory. In particular, identification often does not depend on unverifiable parametric distributional assumptions. This is important: although economics can determine or at least shape the specification of many components of an empirical model, it rarely provides guidance on distribution functions governing unobservables.

[1] A seminal paper in this literature is Paarsch (1992a), which builds on insights in Smiley (1979) and Thiel (1988).

Our focus in this chapter is on structural econometric approaches to auctions, with an emphasis on nonparametric identification. This focus should not be confused with a presumption that nonparametric *estimation* methods are always preferred. Approximation methods are virtually always needed for estimation in finite samples, and parametric estimators will be most appropriate in some applications. However, as emphasized at least since Koopmans (1945), the question of what the observables and the assumed underlying structure are capable of revealing (i.e., the identification question) is fundamentally distinct from the choice of statistical methods used in practice for estimation. When identification holds nonparametrically, one can be confident that estimates have valid interpretations as finite sample approximations and are not merely artifacts of unverifiable maintained assumptions. Equally important for our purpose, because a discussion of nonparametric identification makes clear how the structure of a model and the observables enable (or, in some important cases, fail to enable) estimation, it also provides an ideal perspective for discussing recent developments in empirical approaches to auctions. Our goals are to describe key insights from a wide range of recent work in this area in a unified framework, to present several new results, and to point out areas ripe for exploration.

We focus on two auction formats that are dominant in practice: first-price sealed-bid auctions and ascending (or "English") auctions. First-price auctions are particularly prevalent in government procurement – a common source of data in applied work. Our discussion of first-price auctions will include the closely related Dutch auction. Ascending auctions, in several variations, are the most frequently observed in practice. They are widely used in sales of antiques, art, timber, and in Internet auctions.[2] As we will see below, each type of auction presents different econometric challenges. We will also examine extensions to other environments, including multi-unit and multi-object auctions. We consider a wide range of underlying structures of bidder demand and information, as well as a range of types of data one may encounter in applications. We discuss identification, testable restrictions, and a variety of parametric, semi-parametric, and nonparametric estimation approaches. Much of the recent innovation in the literature has been on the identification question. In many cases this is because standard statistical methods can be applied for estimation and testing once identification results are obtained. This is not always the case, however, and in some cases the development of methods for estimation has lagged development of identification results. Here and elsewhere, our discussion will point to a number of opportunities for additional work.

[2] Among the auction forms commonly discussed in the theoretical literature, we exclude the second-price sealed-bid ("Vickrey") auction, which is closely related to the ascending auction but uncommon in practice. Some Internet auctions, like those on the eBay site, use a system of proxy bidding that has the flavor of a second-price sealed bid auction, although in practice bidders usually have the ability to observe and respond to the bids of at least some of their opponents, as in an ascending auction (see Lucking-Reiley (2000) for more stylized facts about Internet auctions). Alternative models of Internet auctions are offered by Bajari and Hortaçsu (2003a), Ockenfels and Roth (2006) and Song (2003). We will discuss a structural empirical model based on the last of these in Section 6.3.4.

Before proceeding, we first make precise what is meant by identification in this context. Let $\mathbb{G}$ denote the set of all joint distributions over a specified set of observable random variables. Define a *model* as a pair $(\mathbb{F}, \Gamma)$, where $\mathbb{F}$ is a set of joint distributions over a specified set of latent random variables and Γ is a collection of mappings $\gamma : \mathbb{F} \to \mathbb{G}$. In this chapter, $\mathbb{F}$ will typically be the set of joint distributions of bidder valuations and information ("types") satisfying certain statistical properties (e.g., independence, symmetry, etc.), while Γ will consist of a single mapping from the true distribution of types to a distribution of bids implied by the assumption of Bayesian Nash equilibrium. Implicit in the specification of a model is the assumption that it contains the true structure $(\mathcal{F}, \gamma)$ generating the observables. A model is said to be identified (or identifiable) if the observables uniquely determine the true structure within $(\mathbb{F}, \Gamma)$.

DEFINITION 1.1. A model $(\mathbb{F}, \Gamma)$ is *identified* if for every $(\mathcal{F}, \widetilde{\mathcal{F}}) \in \mathbb{F}^2$ and $(\gamma, \tilde{\gamma}) \in \Gamma^2$, $\gamma(\mathcal{F}) = \tilde{\gamma}(\widetilde{\mathcal{F}})$ implies $(\mathcal{F}, \gamma) = (\widetilde{\mathcal{F}}, \tilde{\gamma})$.

In some cases, useful inferences can be made even when a model is not identified. In *partially identified* models one may be able to identify some components of interest, or place bounds on components of interest [see, e.g., Manski (1995)]. A separate question is whether a model places refutable restrictions on observables; i.e., whether the model is testable. A model is testable if some joint distributions in $\mathbb{G}$ cannot be generated by the model.

DEFINITION 1.2. A model $(\mathbb{F}, \Gamma)$ is *testable* if $\bigcup_{\gamma \in \Gamma, \mathcal{F} \in \mathbb{F}} \gamma(\mathcal{F})$ is a strict subset of $\mathbb{G}$.

With these definitions in hand, we can preview some of the themes that emerge in what follows. First, a remarkable number of positive nonparametric identification results can be obtained by exploiting the relationships between observables and the primitives of interest that are implied by economic theory. Richer statistical structures (e.g., arbitrary correlation) for bidders' information and/or more limited sets of observables (e.g., only the winning bid) create greater challenges, but even here a number of positive results can be obtained. There are limits to the positive results, however. For example, identification of models with common values, risk aversion, or unobserved heterogeneity can be obtained only with strong *a priori* restrictions. This is particularly the case in ascending auctions, where theory provides less guidance on the appropriate interpretation of the observed bids.

A second major theme is the need to make modeling choices and the importance of testing these choices when possible. Often a particular set of assumptions (e.g., independent private values, risk neutral bidders) is postulated for particular application based on characteristics of the relevant market. Modeling choices can have important implications for the conclusions one reaches. Ideally, researchers would like to combine economic justifications for modeling choices with statistical evidence supporting these choices and/or an analysis of the range of outcomes possible under alternative

assumptions. We discuss a number of results that clarify when this will be possible. In many cases some assumptions can be tested while maintaining others. In general this possibility depends on the auction format (e.g., ascending versus first-price) and the data configuration (e.g., whether all bids are observed or just the winning bid, or whether particular types of exogenous variation are present and observable). For example, an assumption of private values is testable under some data configurations but not others (Section 8). Another example arises when the econometrician must make modeling choice regarding the source of observed correlation among bids: this may result from correlation of bidders' private information or from common knowledge among bidders of auction-specific factors affecting all bidders' valuations. These alternative models often have different implications for counterfactuals, but it is difficult to distinguish between them empirically (Section 6.1.2). Other examples include choices of how participation is modeled (Section 6.3), and of bidders' risk preferences (Section 6.4). Typically, some modeling choices will be testable (particularly in data configurations that include some type of exogenous variation in the environment – e.g., in the number of bidders or bidder covariates), while others will not.

Our focus in this chapter unavoidably leads us to ignore many interesting and important issues given attention in the empirical auctions literature. Fortunately, there are now several excellent surveys, each with a somewhat different focus, that provide useful complements to our chapter. Laffont (1997) and Hendricks and Paarsch (1995) provide early surveys reviewing empirical studies of the implications of equilibrium bidding in auctions as well as approaches to estimation of the primitives of auction models. Perrigne and Vuong (1999) survey methods for structural analysis of first-price auctions, including a synthesis of their own extensive contributions (with several coauthors) to nonparametric identification and estimation of these models. Hong and Shum (2000) provide an introduction to parametric structural approaches. Kagel (1995) surveys the extensive work on auctions in the experimental economics literature. Finally, Hendricks and Porter (in press) provide a recent and extensive review of the large empirical literature on auctions, covering a wide range of economic questions and econometric approaches.[3]

The structure of the chapter can be described as follows. Section 2 describes the underlying theoretical framework for our initial focus and provides the characterizations of equilibrium bidding behavior that underlie the econometric approaches that follow.[4] In Sections 3 and 4 we then discuss first-price and ascending auctions in the simplest and most widely considered case of private values, assuming that there is no binding

[3] Reiss and Wolak (Chapter 64 in this volume) include a discussion of auctions among several examples of the structural empirical approach in industrial organization. See also the recent monograph by Hong and Paarsch (2006).

[4] For additional detail, Krishna (2002) provides an excellent synthesis of a large theoretical literature on auctions. McAfee and McMillan (1987) provide a shorter introduction to much of the relevant theory. Milgrom and Weber (1982) is a central paper in the early theoretical literature that covers many of the models we consider. Milgrom (2004) treats some of the newer literature on combinatorial auctions.

reserve price and that the data available consist of bids from independent auctions of identical goods. These results provide many of the key building blocks for considering richer private values specifications, specification testing, endogenous participation, risk aversion, as well as other types of data in Sections 5 and 6. In Section 7 we take up the case of common values models, where identification is more difficult and often fails. This provides one motivation for a discussion of testing for common values in Section 8. We conclude with two sections on important topics that have been the subject of very recent work. Section 9 addresses dynamics. Section 10 discusses work in progress on multi-unit and multi-object auctions.

2. Theoretical framework

2.1. Demand and information structures

Throughout we denote random variables in upper case and their realizations in lower case. We use boldface to indicate vectors. To emphasize the distinction between latent variables and observables, we adopt the convention of denoting the cumulative distribution function (CDF) of a latent random variable $\mathbf{Y}$ by $F_{\mathbf{Y}}(\cdot)$ and the CDF of an observable random variable $\mathbf{Y}$ by $G_{\mathbf{Y}}(\cdot)$. Much of the discussion will involve order statistics. We let $Y^{(k:n)}$ denote the kth order statistic from the sample $(Y_1, \ldots Y_n)$, with $F_Y^{(k:n)}(\cdot)$ denoting the corresponding marginal CDF. We follow the standard convention of indexing order statistics lowest to highest so that, e.g., $Y^{(n:n)}$ is the maximum.

For most of the chapter, the underlying theoretical framework involves the sale of a single indivisible good to one of $n \in \{\underline{n}, \ldots, \overline{n}\}$ risk neutral bidders, with $\overline{n} \geqslant \underline{n} \geqslant 2$.[5] Later, when we consider auctions with reserve prices or participation costs, these n bidders will be referred to as "potential bidders." We consider risk aversion, sequential auctions, multi-unit auctions, and multi-object auctions separately below. We let $\mathcal{N}$ denote the set of bidders, although when bidders are symmetric, $n = |\mathcal{N}|$ will be a sufficient statistic. We let $\mathcal{N}_{-i}$ denote the set of competitors faced by bidder i. The utility bidder $i \in \{1, \ldots, \overline{n}\}$ would receive from the good is U_i, which we assume to have common support (denoted supp $F_{U_i}(\cdot)$ or supp U_i) for all i. Often U_i is referred to as i's "valuation." We let $\mathbf{U} = (U_1, \ldots, U_n)$.

Bidder i's private information (his "type") consists of a scalar signal X_i. We let $\mathbf{X} = (X_1, \ldots, X_n)$, $\underline{x}_i = \inf \text{supp}\, X_i$, and $\overline{x}_i = \sup \text{supp}\, X_i$. Signals are informative in the sense that the expectation

$$E[U_i \mid X_i = x_i,\ \mathbf{X}_{-i} = \mathbf{x}_{-i}]$$

strictly increases in x_i for all realizations $\mathbf{x}_{-i}$ of i's opponents' signals. Note that because signals play a purely informational role and any monotonic transformation $\theta(X_i)$

[5] Translation to procurement settings, where bidders compete to sell, is straightforward.

contains the same information as X_i itself, the marginal distribution of X_i is irrelevant; i.e., without a normalization on X_i, the theoretical model is over-parameterized. It is therefore desirable (and without loss of generality) to impose a normalization such as[6]

$$X_i = E[U_i \mid X_i].$$

We will see below that different normalizations will sometimes turn out be more convenient.

Except where otherwise stated, we assume that the set of bidders and the joint distribution $F_{\mathbf{X},\mathbf{U}}(\cdot;\mathcal{N})$ of bidders' signals and valuations are common knowledge. While these are standard assumptions in the theoretical literature on auctions, in a few cases (e.g., an ascending auction with private values) these assumptions are inconsequential. In a first-price auction, these assumptions can be relaxed somewhat; for example, we consider the possibility that $\mathcal{N}$ is unknown in Section 6.3.3.

This framework is a generalization of that studied in Milgrom and Weber's (1982) influential theoretical exploration of auctions and nests a wide range of special cases, each involving different assumptions about bidders' private information. One key distinction is that between *private values* (PV) and *common values* (CV) models.

DEFINITION 2.1. Bidders have *private values* if $E[U_i \mid X_1 = x_1, \ldots, X_n = x_n] = E[U_i \mid X_1 = x_1]$ for all $x_1, \ldots, x_n$ and all i; bidders have *common values* if $E[U_i \mid X_1 = x_1, \ldots, X_n = x_n]$ strictly increases in x_j for all i, j, and x_j.[7]

In private values models, bidders do not have private information about the valuations of their opponents. For the settings we will consider, this is equivalent to assuming bidders know their own valuations ($X_i = U_i$). In a common values model, by contrast, each bidder i would update her beliefs about her valuation U_i if she learned an opponent's signal X_j in addition to her own signal X_i. Even in a private values auction a bidder would like to know her competitors' private information for strategic reasons. However, in a common values auction, knowledge of opponents' signals would alter her expectation of her own valuation. This is the characteristic of common values auctions that leads to the "winner's curse." Roughly speaking, winning a common values auction reveals (in equilibrium) to the winner that her signal was more optimistic than those of her opponents. Rational bidders anticipate this information when forming expectations

[6] It is important to avoid confusing this extra degree of freedom in the usual specification of the theoretical model with issues concerning econometric identification. Since the marginal distribution of X_i is irrelevant in the theoretical model, it is not a primitive whose identification should even be considered.

[7] Alternatively, one might define private and common values in terms of the conditional distributions $F_{U_i}(U_i \mid X_1, \ldots, X_n)$ and $F_{U_i}(U_i \mid X_i)$. For our purposes a definition in terms of conditional expectations is adequate. Note that for simplicity of exposition our definition of common values rules out cases where the winner's curse arises for some realizations of types but not others.

of the utility they would receive by winning.[8] Note that common values models incorporate a wide range of structures in which information about the value of the good is dispersed among bidders, not just the special case in which the value of the object is identical for all bidders (defined as *pure common values* below).[9]

A second way in which this general framework can be specialized is through restrictions on the joint distribution of signals. Common assumptions are independence or affiliation.[10] Note that dependence (or affiliation) of signals is neither necessary nor sufficient for common values. Finally, a common restriction in the literature is *symmetry*, i.e., that the joint distribution $F_{\mathbf{X},\mathbf{U}}(X_1, \ldots, X_n, U_1, \ldots, U_n; \mathcal{N})$ is exchangeable in the bidder indices. For clarity, we will often explicitly refer to models as "symmetric" or "asymmetric." Combining these types of restrictions leads to a number of special cases that have been considered in the literature, including:

- *Independent Private Values (IPV)*: private values with U_i independent;
- *Symmetric Independent Private Values*: private values with U_i i.i.d.;
- *Affiliated Private Values (APV):* private values with $(U_1, \ldots, U_n)$ affiliated;
- *Pure Common Values*: common values with $U_i = U_0 \ \forall i$;
- *Mineral Rights*: pure common values with signals i.i.d. conditional on U_0.

Finally, for a few results we will make an additional assumption of *exogenous variation in the number of bidders*, which holds when variation in the set of bidders is independent of the joint distribution of bidders' valuations and signals.

DEFINITION 2.2. A bidding environment has *exogenous variation in the number of bidders* if $\overline{n} > \underline{n}$ and, for all $\mathcal{N}, \mathcal{N}'$ such that $\mathcal{N} \subset \mathcal{N}' \subseteq \{1, \ldots, \overline{n}\}$, $F_{\mathbf{X},\mathbf{U}}(\cdot; \mathcal{N})$ is identical to the marginal distribution of $\{(U_i, X_i)\}_{i \in \mathcal{N}}$ obtained from $F_{\mathbf{X},\mathbf{U}}(\cdot; \mathcal{N}')$.

2.2. Equilibrium bidding

We restrict attention to econometric approaches that exploit the structure of equilibrium bidding to obtain identification or testable restrictions. Hence we must first provide the

[8] Note that the presence of the winner's curse does not imply that winners regret winning; rather, the winner's curse refers to the "bad news" [Milgrom (1981)] about the object's value contained the information that one has won the auction. Rational bidders anticipate this.

[9] While our terminology follows, e.g., Klemperer (1999), Athey and Haile (2002), and Haile, Hong and Shum (2003), there is some variation in the terminology used in the auction literature. Early on, the term "common values" was sometimes used in the way we use it but sometimes used to refer to the special case we call "pure common values." Similarly, "affiliated values" was sometimes used for the class of models we call "common values," despite the fact that purely private values can be affiliated (see below). Recently some authors [e.g., Krishna (2002)] have adopted the term "interdependent values" to refer to the broad class of models we refer to as common values models.

[10] The random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$ with joint density $f_{\mathbf{Y}}(\cdot)$ are affiliated if for all $\mathbf{y}$ and $\mathbf{y}'$, $f_{\mathbf{Y}}(\mathbf{y} \vee \mathbf{y}') f_{\mathbf{Y}}(\mathbf{y} \wedge \mathbf{y}') \geqslant f_{\mathbf{Y}}(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}')$, where $\vee$ denotes the component-wise maximum, and $\wedge$ the component-wise minimum. See Milgrom and Weber (1982) for additional discussion. Note that affiliation allows independence as a special case.

necessary characterizations of equilibrium. Following the literature, we generally restrict attention to (perfect) Bayesian Nash equilibria in weakly undominated strategies. We focus on equilibrium in pure bidding strategies $\beta_i(\cdot; \mathcal{N})$, $i = 1, \ldots, n$, mapping each bidder's signal (and, implicitly, any public information) into a bid. When bidders are *ex ante* symmetric we further restrict attention to symmetric equilibria, so that $\beta_i(\cdot) = \beta(\cdot)\ \forall i$. Below we discuss conditions under which there are other equilibria in first-price auctions. We will denote a bidder i's equilibrium bid by B_i, with $\mathbf{B} = \{B_1, \ldots, B_n\}$. We let $\underline{b}_i = \inf[\text{supp}[B_i]]$ and $\overline{b}_i = \sup[\text{supp}[B_i]]$.

2.2.1. First-price auctions

In a first-price sealed-bid auction bidders submit bids simultaneously, and the good is awarded to the high bidder at a price equal to his bid. If there is a reserve price, r, the seller has committed to consider only bids of at least r. For first-price auctions we make the following additional assumptions:

ASSUMPTION 2.1 *(First-price auction assumptions).*

(i) For all i, U_i has compact, convex support denoted $\text{supp}\, F_{U_i}(\cdot) = [\underline{u}, \overline{u}]$.
(ii) The signals $\mathbf{X}$ are affiliated, with $\text{supp}\, F_{\mathbf{X}}(\cdot) = \times_{i=1}^{n} \text{supp}\, F_{X_i}(\cdot)$.
(iii) $F_{\mathbf{X}}(\cdot)$ has an associated joint density $f_{\mathbf{X}}(\cdot)$ that is strictly positive on the interior of $\text{supp}\, F_{\mathbf{X}}(\cdot)$.

The following result summarizes existence and uniqueness results for this model. This will enable us to then proceed to the key characterization results used for empirical work.

THEOREM 2.1. *Consider the first-price auction.*

(i) (Existence in strictly increasing strategies) *An equilibrium exists in pure, nondecreasing strategies, where for each* i, $\text{supp}[B_i] \subseteq \text{supp}[\max_{j \in \mathcal{N} \setminus i} B_j]$. *In addition, a pure strategy equilibrium in strictly increasing strategies exists in all models except in the CV model with asymmetric bidders and signals that are not independent*; *in the latter case, strategies are strictly increasing except that at most one bidder may bid* $\inf[\text{supp}[B^{(n:n)}]]$ *with positive probability.*[11]
(ii) (Uniqueness) *In a PV model with either* (a) *independence* (*IPV*), *or* (b) *symmetry, if* $f_{\mathbf{X}}(\cdot)$ *is continuously differentiable there is a unique equilibrium. This equilibrium is in pure, strictly increasing, and differentiable strategies.*[12]

[11] See Athey (2001) and Reny and Zamir (2004) for existence of equilibrium in nondecreasing strategies, and Milgrom and Weber (1982), McAdams (2007) and Lizzeri and Persico (2000) for the characterization. McAdams (2007) argues that in any monotone equilibrium, strategies are strictly increasing except that at most one bidder may, with positive probability, choose the lowest bid that wins with strictly positive probability, if such a bid exists. In PV auctions it is possible to rule out mass points at the reserve price, if it binds, or at the bottom of the value distribution if the reserve price does not bind.

[12] In the IPV case, all equilibria are in monotone strategies; see Lebrun (1999), Bajari (2001), and Maskin and Riley (2003) for uniqueness results. Milgrom and Weber (1982) show existence of the equilibrium for

(iii) (Uniqueness in monotone class for symmetric models) *If we restrict attention to pure strategy equilibria in nondecreasing strategies, then when bidders are symmetric and $f_{\mathbf{X}}(\cdot)$ is continuously differentiable there is a unique equilibrium, which is in symmetric, strictly increasing, and differentiable strategies.*[13]

All of our positive identification results for common values models rely on symmetry, so in our discussions of CV auctions we will proceed under the assumption that strategies are strictly increasing. For the first-price auction models for which uniqueness has not been established, we will also assume that all observations in a given data set are derived from the same equilibrium.

As shown by Milgrom and Weber (1982), a bidder i participates if and only if his signal exceeds a threshold value

$$x_i^*(\mathcal{N}) = \inf\left\{x_i\colon E\left[U_i \;\middle|\; X_i = x_i,\ \max_{j\in\mathcal{N}_{-i}} B_j \leqslant r\right] \geqslant r\right\}. \tag{2.1}$$

When there is no reserve price, let $x_i^*(\mathcal{N}) = \underline{x}_i$. Here, expectations over others' bids represent equilibrium expectations. A bidder i who has observed signal $X_i = x_i > x_i^*(\mathcal{N})$ solves

$$\max_{\tilde{b}}\left(E\left[U_i \;\middle|\; X_i = x_i,\ \max_{j\in\mathcal{N}_{-i}} B_j \leqslant \tilde{b}\right] - \tilde{b}\right)\Pr\left(\max_{j\in\mathcal{N}_{-i}} B_j \leqslant \tilde{b} \;\middle|\; X_i = x_i\right), \tag{2.2}$$

where we adopt the convention that $B_j < r$ for any bidder j who does not participate.

Define

$$\tilde{v}_i(x_i, m_i; \mathcal{N}) = E\left[U_i \;\middle|\; X_i = x_i,\ \max_{j\in\mathcal{N}_{-i}} B_j = m_i\right].$$

This is bidder i's expectation of his valuation conditional on his own signal and the highest competing bid. This highest competing bid is informative because, in equilibrium, bids are strictly increasing in signals. In particular, if we let

$$v_i(x_i, y_i; \mathcal{N}) = E\left[U_i \;\middle|\; X_i = x_i,\ \max_{j\in\mathcal{N}_{-i}} B_j = \beta_i(y_i; \mathcal{N})\right] \tag{2.3}$$

APV auctions. McAdams (2007) shows that for a nonmonotone equilibrium to exist, both independence of signals and private values must be relaxed. He shows that with private values or independent signals, all equilibria are outcome-equivalent to a monotone equilibrium; i.e., bidding strategies are identical to those in a monotone equilibrium except possibly for subsets of types whose equilibrium bids win with probability zero. McAdams (2004b) shows that if bidders are symmetric, there is a unique equilibrium within the monotone class. So together, these results imply that for the symmetric PV model, there is a unique equilibrium.

[13] See footnote 12 for a discussion of when nonmonotone equilibria can exist. McAdams (2004b) proves uniqueness within the monotone class. For characterizations, see Milgrom and Weber (1982). See also Lizzeri and Persico (2000), who show that when the density of the value distribution is C^1, in two-bidder first-price auctions with a binding reserve price, among monotone pure strategy equilibria there exists a unique equilibrium in strictly increasing, differentiable strategies, except that one bidder may choose the reserve price with positive probability.

then

$$v_i(x_i, y_i; \mathcal{N}) = \tilde{v}_i\big(x_i, \beta_i(y_i; \mathcal{N}); \mathcal{N}\big).$$

The expectation $v_i(x_i, x_i; \mathcal{N})$ will play an important role below. This expectation is taken conditioning both on i's own private information and on the event that i's equilibrium bid is "pivotal," i.e., that infinitesimal deviations from his equilibrium bid would change the outcome of the auction.

Let

$$G_{M_i|B_i}(m_i|b_i; \mathcal{N}) = \Pr\Big(\max_{j\neq i} B_j \leqslant m_i \;\Big|\; B_i = b_i, \mathcal{N}\Big)$$

denote the distribution of the maximum equilibrium bid among i's opponents conditional on i's own equilibrium bid and the set of bidders $\mathcal{N}$. Let $g_{M_i|B_i}(m_i|b_i; \mathcal{N})$ denote the corresponding conditional density, which exists and is positive for all b_i and almost every m_i in the support of B_i under the assumptions outlined above. Note that with strictly increasing equilibrium bidding, conditioning on $\{B_i = b\}$ is equivalent to conditioning on $\{X_i = \beta_i^{-1}(b; \mathcal{N})\}$. Bidder i's bidding Problem (2.2) can then be rewritten

$$\max_{\tilde{b}} \int_{-\infty}^{\tilde{b}} \big[\tilde{v}_i(x_i, m_i; \mathcal{N}) - \tilde{b}\big] g_{M_i|B_i}\big(m_i|\beta_i(x_i; \mathcal{N}); \mathcal{N}\big)\, \mathrm{d}m_i.$$

This objective function is differentiable almost everywhere. Differentiating with respect to $\tilde{b}$, we see that for almost every signal x_i of bidder i, a necessary condition for b_i to be an optimal bid (i.e., for $\beta_i(x_i; \mathcal{N}) = b_i$) is

$$v_i(x_i, x_i; \mathcal{N}) = b_i + \frac{G_{M_i|B_i}(b_i|b_i; \mathcal{N})}{g_{M_i|B_i}(b_i|b_i; \mathcal{N})} \equiv \xi_i(b_i; \mathcal{N}). \tag{2.4}$$

Equation (2.4) characterizes an equilibrium bid as equal to the bidder's expectation of his valuation (conditional on being pivotal) less a strategic "markdown" $\frac{G_{M_i|B_i}(b_i|b_i;\mathcal{N})}{g_{M_i|B_i}(b_i|b_i;\mathcal{N})}$.[14] This first-order condition does not always lead to an analytic solution for equilibrium bidding strategies. With *ex ante* symmetric bidders, however, we can write

$$v_i(x, x; \mathcal{N}) = v(x, x; n) = E\Big[U_i \;\Big|\; X_i = \max_{j\neq i} X_j = x\Big]$$

and $x_i^*(\mathcal{N}) = x^*(n)\ \forall i$. In that case, Milgrom and Weber (1982) have shown that the equilibrium bid function has the form

$$\beta(x; n) = rL\big(x^*(n)|x; n\big) + \int_{x^*(n)}^{x} v(t, t; n)\, \mathrm{d}L(t|x; n) \tag{2.5}$$

[14] This is analogous to the markdown of an oligopsonist, which bases its price on the equilibrium elasticity of its residual supply curve; in the auction model, $G_{M_i|B_i}(b_i|b_i; \mathcal{N})$ plays the role of the residual supply curve.

for $x \geqslant x^*$, where

$$L(t|x;n) \equiv \exp\left(-\int_t^x \frac{f_1(z|z;n)}{F_1(z|z;n)}\,\mathrm{d}z\right)$$

and $F_1(\cdot|x;n)$ is the distribution of the maximum signal among a bidder's opponents conditional on the number of bidders and on his own signal being x.

Before proceeding, we pause to make two observations about the support of the equilibrium bid distribution.[15]

THEOREM 2.2. *In the IPV model of the first-price auction,* $\mathrm{supp}[B_i]$ *is the same for all* i.

PROOF. With independence, the inverse bid function for bidder i can be written

$$\xi_i(b_i;\mathcal{N}) = b_i + \frac{1}{\sum_{k\in\mathcal{N}\setminus i}\frac{g_{B_k}(b_i)}{G_{B_k}(b_i)}}.$$

If there are two bidders, the result is immediate given that the value distributions have the same support. Now suppose $n > 2$ and $\underline{b}_i < \underline{b}_j$. We know that $\xi_j(b_j;\mathcal{N})$ must be continuous at $\underline{b}_i$: otherwise (given strictly monotone strategies) we would contradict our assumption that valuations are drawn from a convex set. Then, note that

$$\underline{u} = \xi_i(\underline{b}_i;\mathcal{N}) = \underline{b}_i + \frac{1}{\sum_{k\in\mathcal{N}\setminus i}\frac{g_{B_k}(\underline{b}_i)}{G_{B_k}(\underline{b}_i)}} < \underline{b}_i + \frac{1}{\sum_{k\in\mathcal{N}\setminus\{i,j\}}\frac{g_{B_k}(\underline{b}_i)}{G_{B_k}(\underline{b}_i)}} = \xi_j(\underline{b}_i;\mathcal{N}).$$

But $\xi_j(\underline{b}_i;\mathcal{N}) > \underline{u}$ contradicts the assumption that U_i has the same support for all i. Given the properties established in Theorem 2.1, standard arguments then show that $\mathrm{supp}[B_i] = [\max\{r, \underline{u}\}, \bar{b}]\ \forall i$. □

Outside of the IPV model, it is not known in general whether bid distributions have the same support for all bidders when bidders are asymmetric. We do know that if we relax the assumption that valuations have common support, the bids may or may not have the same support.[16]

Note that the theory also implies that the upper bound of the bid distribution is closely related to features of the distribution of valuations. In the symmetric IPV model,

$$U_i = B_i + \frac{G_B(B_i;n)}{(n-1)g_B(B_i;n)}$$

[15] See Lebrun (1999) for an alternative proof.

[16] In Section 5.1 we give an example where valuations have different supports but bids have identical supports. To see an example where bid distributions have different supports, suppose that there are three bidders. $F_{U_1}(u_1) = \frac{8}{5}u_1 - \frac{16}{25}u_1^2$ for $u_1 \in [0, 5/4]$, while for $i \in \{2,3\}$, $F_{U_i}(u_i) = \frac{1}{100}(4 + 2u_i - \sqrt{2}\sqrt{8 - 7u_i + 2u_i^2})^2$ for $u_i \in [0, 3/2]$ and $F_{U_i}(u_i) = \frac{1}{9}u_i^2$ for $u_i \in [3/2, 3]$. For this example, $G_{B_1}(b_1) = 2b_1 - b_1^2$ for $b_1 \in [0,1]$, while for $i \in \{2,3\}$, $G_{B_i}(b_i) = b_i^2/4$ for $b_i \in [0,2]$.

so

$$E[U_i] = E[B_i] + \frac{1}{n-1}\int_{\underline{b}}^{\overline{b}} G_B(b;n)\,\mathrm{d}b = \frac{n-2}{n-1}E[B_i] + \frac{1}{n-1}\overline{b}.$$

Thus, the mean valuation is a linear function of the mean bid and $\overline{b}$. When $n = 2$, this yields $E[U_i] = \overline{b}$. The average "markdown" for a bidder in the symmetric IPV model is given by

$$E[U_i - B_i] = \frac{1}{n-1}\big(\overline{b} - E[B_i]\big).$$

Although it seems that these kinds of relationships might be useful, they have not to our knowledge been explored in the econometric analysis of auctions.

2.2.2. Ascending auctions

The standard model of an ascending auction is the so-called "clock auction" or "button auction" model of Milgrom and Weber (1982), where the price rises continuously and exogenously. Bidders indicate their willingness to continue bidding continuously as well, for example by raising their hands or depressing a button as the price rises. As the auction proceeds, bidders exit observably and irreversibly (by lowering their hands, releasing their buttons, etc.) until only one bidder remains. This final bidder obtains the good at the price at which his last opponent exited; i.e., the auction ends at a price equal to the second highest exit price ("bid") $b^{(n-1:n)}$.

The participation rule for an ascending auction is identical to that for a first-price auction. An equilibrium bidding strategy specifies a price at which to exit, conditional on one's own signal and on any information revealed by previous exits by opponents. With strictly increasing bidding strategies, the price at which a bidder exits reveals his signal to others. So in a common values auction, an exit causes the remaining bidders to update their beliefs about their valuations; hence, the prices at which bidders plan to exit change as the auction proceeds. In a private values auction there is no such updating, and each bidder has a weakly dominant strategy to bid up to his valuation, i.e.,

$$\beta_i(x_i;\mathcal{N}) = E[U_i \mid X_i = x_i] = x_i \equiv u_i. \tag{2.6}$$

In common values auctions there are multiple equilibria, even with *ex ante* symmetric bidders and restriction to symmetric strictly increasing weakly undominated strategies [Bikhchandani, Haile and Riley (2002)]. In any such equilibrium, however, if i is one of the last two bidders to exit, his exit price b_i is

$$E\big[U_i \;\big|\; X_i = x_i,\ X_j = x_i\ \forall j \notin \{i \cup \mathcal{E}_i\},\ X_k = x_k\ \forall k \in \mathcal{E}_i\big], \tag{2.7}$$

where $\mathcal{E}_i$ denotes the set of bidders who exit before i. Milgrom and Weber (1982) originally identified the equilibrium in which all bidders follow (2.7), which reduces to the weakly dominant strategy (2.6) in the case of private values.

While the Milgrom–Weber model yields a trivial relation between a bidder's valuation and his bid in a private values auction, we will see that even in this case identification can present challenges, due to the fact that the auction ends before the winner bids (exits). Furthermore, in many applications the Milgrom–Weber model may represent too great an abstraction from actual practice, for example if prices are called out by bidders rather than by the auctioneer, or if bidders are free to make a bid at any point in the auction, regardless of their activity (or lack thereof) earlier in the auction. In Section 4.3 we will discuss an econometric approach that relaxes the structure of the button auction model.

3. First-price auctions with private values: Basic results

3.1. Identification

We begin by considering the case of private values auctions, assuming that bidders' valuations at each auction are draws from the same joint distribution $F_{\mathbf{U}}(\cdot)$. The primitive of interest in a PV auction is this joint distribution: it completely characterizes bidder demand and information. With knowledge of $F_{\mathbf{U}}(\cdot)$ one can, for example, simulate outcomes under alternative market mechanisms, assess efficiency and the division of surplus, and determine an optimal reserve price. The simple idea underlying the structural approach to PV auctions is to use the distribution of bids observed in a sample of auctions along with the equilibrium mapping between valuations and bids (the observables) to learn about $F_{\mathbf{U}}(\cdot)$.

Even when a closed form solution like (2.5) is available, however, it is not immediately clear how one would proceed to use this equilibrium characterization for a first-price auction to obtain identification. Even in the simplest symmetric IPV model, the equilibrium bid function takes the form (recall that $x_i = u_i$)

$$\beta(u; n) = \frac{\int_{-\infty}^{u} t f_U^{(n-1:n-1)}(t)\,\mathrm{d}t}{F_U^{(n-1:n-1)}(u)},$$

which depends on the unknown distribution $F_U(\cdot)$ of valuations, i.e., on the object one would like to estimate.

Several approaches were initially taken to address this problem within the symmetric IPV model. Following Smiley (1979) and Paarsch (1992a), early work focused on parametric specifications of $F_U(\cdot)$ admitting simple closed form equilibrium bid functions that made it feasible to derive likelihoods or moment conditions.[17] Laffont, Ossard and Vuong (1995) proposed an approach combining parametric assumptions with a simulation based estimator that is made feasible in the symmetric IPV framework by the revenue equivalence theorem [e.g., Myerson (1981)]. Bajari (1997) proposed a Bayesian

[17] Smiley (1979) considered only common values models.

approach applicable in the more difficult case of asymmetric independent private values. The role of the parametric distributional assumptions in these empirical approaches was not initially clear.

An important breakthrough due to Guerre, Perrigne and Vuong (2000) came from the simple but powerful observation that equilibrium is attained when each player is acting optimally against the distribution of behavior by opponents.[18] When bids are observable, both the distribution of opponent behavior and the optimal (equilibrium) action of each bidder are observable, enabling identification of the latent joint distribution of bidder valuations under fairly weak restrictions. In particular, the first-order condition (2.4) can be written

$$u_i = b_i + \frac{G_{M_i|B_i}(b_i|b_i;\mathcal{N})}{g_{M_i|B_i}(b_i|b_i;\mathcal{N})}. \tag{3.1}$$

Thus, each bidder's latent private value can be expressed as a functional of his equilibrium bid and the joint distribution of the competing equilibrium bids he faces.[19] In fact, the function $\xi_i(b_i,\mathcal{N}) \equiv b_i + \frac{G_{M_i|B_i}(b_i|b_i;\mathcal{N})}{g_{M_i|B_i}(b_i|b_i;\mathcal{N})}$ is the inverse of bidder i's equilibrium bid function, the mapping needed to infer valuations from bids. Since the joint distribution of bids is observable, identification of each private value u_i (and, therefore, of the joint distribution $F_{\mathbf{U}}(\cdot)$) follows directly from (3.1). Formally,

$$F_{\mathbf{U}}(\mathbf{u}) = G_{\mathbf{B}}\big(\xi_1^{-1}(u_1,\mathcal{N}),\ldots,\xi_n^{-1}(u_n,\mathcal{N})\big). \tag{3.2}$$

This proves the following identification result, combining results from Guerre, Perrigne and Vuong (2000), Li, Perrigne and Vuong (2002), and Campo, Perrigne and Vuong (2003).

THEOREM 3.1.

(i) *Suppose all bids are observed in first-price sealed-bid auctions. Then the symmetric affiliated private values model is identified.*

(ii) *Suppose all bids and bidder identities are observed in first-price sealed-bid auctions. Then the asymmetric affiliated private values model is identified.*

3.2. Estimation

For purposes of estimation, suppose one observes bids from independent auctions $t = 1,\ldots,T$. We will add an auction index t to the notation above as necessary. For

[18] This approach was first described in print by Laffont and Vuong (1993), who attribute the idea to an early draft of Guerre, Perrigne and Vuong (2000).

[19] Note that in general this kind of approach relies on there being a unique equilibrium or on an assumption that the equilibrium selected is the same across observations. Otherwise the observed distribution of opponent bids would be a mixture of those in each equilibrium, and would not match the distribution characterizing a bidder's beliefs in a given auction.

example, b_{it} will denote the realized bid of bidder i at auction t. Let $T_{\mathcal{N}}$ denote the number of auctions in which $\mathcal{N}$ is the set of bidders. We let $T_n = \sum_{\mathcal{N}:\,|\mathcal{N}|=n} T_{\mathcal{N}}$. We assume that for all $n = \underline{n} \ldots \overline{n}$, $T_n \to \infty$ and $T \to \infty$. When we consider asymmetric settings, we consider only sets $\mathcal{N}$ for which $T_{\mathcal{N}} \to \infty$.

A two-step estimation procedure can be employed, closely following the identification result in Theorem 3.1. In the first step, estimates of each $\frac{G_{M_i|B_i}(b_{it}|b_{it};\mathcal{N})}{g_{M_i|B_i}(b_{it}|b_{it};\mathcal{N})}$ are obtained from the observed bids. These estimates are then used with Equation (3.1) to construct estimates of each latent valuation u_i. This pseudo-sample of valuations (often referred to as a sample of "pseudo-values") is then treated as a sample from the true distribution $F_{\mathbf{U}}(\cdot)$, subject to first-stage estimation error.

In principle each step could be parametric or nonparametric. As noted by Perrigne and Vuong (1999), a challenge in a fully parametric method is the need for internal consistency between the parametric families chosen for the distributions of bids and of valuations, since these are related by the equilibrium bid function. This issue would be avoided if only one of the two steps were treated parametrically. Jofre-Bonet and Pesendorfer (2003) and Athey, Levin and Seira (2004) follow this approach, motivated by a desire to include covariates in a parsimonious way.[20] Fully parametric methods based on maximum likelihood or moment conditions (rather than the two-step "indirect" approach discussed here) have been explored by, e.g., Paarsch (1992a, 1992b), Donald and Paarsch (1993, 1996), and Laffont, Ossard and Vuong (1995). In practice the applicability of these methods has been limited to distributional families leading to simple closed forms for equilibrium bid functions and/or to the symmetric independent private values setting. As first explored by Donald and Paarsch (1993), a violation of a standard regularity condition for maximum likelihood estimation arises in a first-price auction, leading to nonstandard asymptotic distributions [see also Donald and Paarsch (1996), Chernozhukov and Hong (2003), and Hirano and Porter (2003)].

Below we describe the fully nonparametric estimators that have thus far been proposed in the literature.[21]

3.2.1. Symmetric bidders

Consider first the case of symmetric bidders, where $G_{M_i|B_i}(b|b;\mathcal{N})$ can be written $G_{M|B}(b|b;n)\ \forall i$. Following Li, Perrigne and Vuong (2002), let

$$G_{M,B}(m, b; n) \equiv G_{M|B}(m|b; n)g_B(b; n)$$

[20] Note, however, that theory predicts that bid distributions should have compact support. To be consistent with theory, an upper bound on the support of the bid distributions should be incorporated in estimation.

[21] Thus far, the literature has focused on kernel estimators. One possible alternative is sieve estimation [e.g., Chen (2007)]. As we discuss below, such an approach might have a practical advantage in environments with observed auction heterogeneity.

and

$$g_{M,B}(m; b; n) \equiv g_{M|B}(m|b; n) g_B(b; n)$$

where $g_B(\cdot)$ is the marginal density of a bidder's equilibrium bid, given the number of bidders n. Note that here we depart from our usual notational convention, since $G_{M,B}(\cdot)$ is not the joint distribution of (M, B) but its derivative with respect to its second argument. Let

$$\widehat{G}_{M,B}(b, b; n) = \frac{1}{nT_n h_G} \sum_{t=1}^{T} \sum_{i=1}^{n} K\left(\frac{b - b_{it}}{h_G}\right) \mathbf{1}\{m_{it} < b, n_t = n\}, \tag{3.3}$$

$$\hat{g}_{M,B}(b, b; n) = \frac{1}{nT_n h_g^2} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{1}\{n_t = n\} K\left(\frac{b - b_{it}}{h_g}, \frac{b - m_{it}}{h_g}\right), \tag{3.4}$$

where M_{it} denotes the maximum of i's opponents' bids at auction t, $K(\cdot)$ is a kernel, and h_G and h_g are appropriately chosen bandwidth sequences. Under standard conditions, $\widehat{G}_{M,B}(b, b; n)$ and $\hat{g}_{M,B}(b, b; n)$ are consistent estimators of $G_{M,B}(b, b; n)$ and $g_{M,B}(b; b; n)$. Noting that

$$\frac{G_{M,B}(b, b; n)}{g_{M,B}(b, b; n)} = \frac{G_{M|B}(b|b; n)}{g_{M|B}(b|b; n)}$$

we see that $\frac{\widehat{G}_{M,B}(b,b;n)}{\hat{g}_{M,B}(b,b;n)}$ is a consistent estimator of $\frac{G_{M|B}(b|b;n)}{g_{M|B}(b|b;n)}$. Equation (3.1) then implies that

$$\hat{u}_{it} \equiv b_{it} + \frac{\widehat{G}_{M,B}(b_{it}, b_{it}; n)}{\hat{g}_{M,B}(b_{it}, b_{it}; n)}$$

is a consistent estimate of the latent valuation u_{it} that generated the observed bid b_{it}.

Naively treating each $\hat{u}_{it}$ as a draw from $F_{\mathbf{U}}(\cdot)$ might suggest a kernel density estimator of the form

$$\hat{f}_{\mathbf{U}}(u_1, \ldots, u_n) = \frac{1}{T_n h_f^n} \sum_{t=1}^{T} K_f\left(\frac{u_1 - \hat{u}_{1t}}{h_f}, \ldots, \frac{u_n - \hat{u}_{nt}}{h_f}\right) \mathbf{1}\{n_t = n\},$$

where $K_f(\cdot)$ is a multivariate kernel and h_f is a bandwidth. Li, Perrigne and Vuong (2002, Proposition 2) show that with bandwidths h_G, h_g, and h_f that vanish at appropriate rates, under standard smoothness conditions $\hat{f}_{\mathbf{U}}(\cdot)$ is in fact a uniformly consistent estimator of $f_{\mathbf{U}}(\cdot)$ on any inner compact subset of its support. The restriction to the region of support away from the boundaries follows from the usual problem of asymptotic bias at the boundaries with kernel estimates.

Li, Perrigne and Vuong (2002, pp. 180–181) suggest triweight kernels (using products of univariate kernels for the multivariate kernels) and a standard rule of trimming the pseudo-values associated with bids within one bandwidth of either boundary of the bid data. The most important practical question is the choice of bandwidth. Guerre,

Perrigne and Vuong (2000) and Li, Perrigne and Vuong (2002) suggest following Silverman's (1986) "rule of thumb." To our knowledge, data driven bandwidth selection procedures have not been explored. Guerre, Perrigne and Vuong (2000) also point out that the assumption of exchangeability can be imposed by averaging $\hat{f}_{\mathbf{U}}(u_1, \ldots, u_n)$ over all permutations of the bidder indices. When there is exogenous variation in the number of bidders, it may be useful to further exploit this restriction by optimally combining information from auctions with different numbers of bidders. As we discuss in more detail below, the overidentifying exchangeability restriction or exogenous variation in participation can also serve as a basis for specification testing.

An important but largely unresolved question is the asymptotic distribution of the estimator $\hat{f}_{\mathbf{U}}(\cdot)$. The challenge is to appropriately account for the estimation error arising from the first-stage estimation of the markdown component of the equilibrium bid functions [Guerre, Perrigne and Vuong (2000)]. Of course, one is often interested in confidence intervals on an estimate of some functional of $f_{\mathbf{U}}(\cdot)$, rather than on $\hat{f}_{\mathbf{U}}(\cdot)$ itself. For example, the goal of the empirical exercise may be to determine optimal selling procedures, to assess efficiency, or to describe how valuations are affected by various factors. For the symmetric case, Haile, Hong and Shum (2003) have shown that the estimates $\hat{u}_{it}$ themselves have asymptotic normal distributions, as do all fixed quantiles (and many other functionals) of their empirical distribution. In practice, a bootstrap procedure has sometimes been applied for inference on these functionals of $F_{\mathbf{U}}(\cdot)$ or others expected to have a normal limiting distribution [e.g., Hendricks, Pinkse and Porter (2003), Haile, Hong and Shum (2003), Krasnokutskaya (2004)]. Outside the IPV model, a block bootstrap is used, reflecting the assumption that auctions are independent, whereas bids may be correlated within an auction. In particular, to construct one bootstrap sample of bids for a given value of n, auction indices s are sampled with replacement from the set $\{t\colon\ n_t = n\}$. All bids from each selected auction s are then included in the bootstrap sample. Haile, Hong and Shum (2003) have also explored the use of subsampling.

In the special case of (symmetric) independent private values, the joint distribution $F_{\mathbf{U}}(\cdot)$ is a product of identical marginal distributions, $F_U(\cdot)$, and the first order condition (3.1) simplifies to

$$u = b + \frac{G_B(b; n)}{(n-1)g_B(b; n)}, \tag{3.5}$$

where $G_B(\cdot; n)$ is the marginal distribution of equilibrium bids in auctions with n bidders, and $g_B(\cdot; n)$ is the associated density. Because $G_B(\cdot; n)$ and $g_B(\cdot; n)$ are univariate functions, this simplifies estimation. Let

$$\widehat{G}_B(b; n) = \frac{1}{nT_n} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{1}\{b_{it} \leqslant b,\ n_t = n\},$$

$$\hat{g}_B(b; n) = \frac{1}{nT_n h_g} \sum_{t=1}^{T} \sum_{i=1}^{n} K\left(\frac{b - b_{it}}{h_g}\right) \mathbf{1}\{n_t = n\},$$

$$\hat{u}_{it} = b_{it} + \frac{\widehat{G}_B(b_{it}; n_t)}{(n_t - 1)\hat{g}_B(b_{it}; n_t)},$$

where $K(\cdot)$ is a kernel (satisfying standard conditions) and h_g is an appropriately chosen bandwidth sequence.[22] Guerre, Perrigne and Vuong (2000) show that with appropriately chosen bandwidth sequence h_f, one then obtains a uniformly consistent estimator of $f_U(\cdot)$ from the kernel density estimator

$$\hat{f}_U(u) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_t}\sum_{i=1}^{n_t}\frac{1}{h_f}K\left(\frac{u - \hat{u}_{it}}{h_f}\right).$$

3.2.2. Asymmetric bidders

Extending the approach above to the case of asymmetric bidders is straightforward, but more data intensive. With symmetric bidders, estimation of the distribution of opposing bids (and the markdown term $\frac{G_{M,B}(b|b;n)}{g_{M,B}(b|b;n)}$ this distribution implies) is performed separately for each value of n. This reflects the fact that variation in n changes the distribution of the maximum opposing bid and, therefore, the equilibrium bidding strategy that is inverted to recover private values from the observed bids. With asymmetric bidders, variation in the *identities* of opposing bidders can have a similar effect, even when the *number* of opponents is held constant. Depending on the nature of bidder asymmetries, different approaches will be taken, although the general principle is clear: to estimate the markdown $\frac{G_{M_i|B_i}(b_i|b_i;\mathcal{N}_t)}{g_{M_i|B_i}(b_i|b_i;\mathcal{N}_t)}$ for a bidder i in auction t, the relevant sample is the set of auctions s in which $G_{M_i|B_i}(\cdot|\cdot;\mathcal{N}_s) = G_{M_i|B_i}(\cdot|\cdot;\mathcal{N}_t)$.

In the most general case, each bidder is allowed to draw her valuation from a different distribution and each set of bidders $\mathcal{N}$ is treated separately. Again let M_{it} denote the maximum bid among i's opponents at auction t. Letting $T_{\mathcal{N}_t}$ denote the number of auctions in which the set of bidders is $\mathcal{N}_t \ni i$, one could let

$$\widehat{G}_{M,B}(b_{it}, b_{it}; \mathcal{N}_t) = \frac{1}{T_{\mathcal{N}_t} h_G}\sum_{s=1}^{T} K\left(\frac{b_{it} - b_{is}}{h_G}\right)\mathbf{1}\{m_{is} < b_{it},\ \mathcal{N}_s = \mathcal{N}_t\},$$

$$\hat{g}_{M,B}(b_{it}, b_{it}; \mathcal{N}_t) = \frac{1}{T_{\mathcal{N}_t} h_g^2}\sum_{s=1}^{T}\mathbf{1}\{\mathcal{N}_s = \mathcal{N}_t\}K\left(\frac{b_{it} - b_{is}}{h_g}\right)K\left(\frac{b_{it} - m_{is}}{h_g}\right)$$

[22] See Guerre, Perrigne and Vuong (2000) for details. They also propose kernel smoothing over the different values of n in estimating each $G_B(\cdot; n)$ and $g(\cdot; n)$ rather than the pure "binning" approach described here. Asymptotically there is no difference and, since N is discrete, kernel smoothing is a generalization. In finite sample, kernel smoothing that is not equivalent to binning will utilize bids from auctions with $n' \neq n$ bidders to estimate the markdown $\frac{G_B(b;n)}{(n-1)g_B(b;n)}$ in (3.5). Whether this is desirable will depend on the data available, although we are not aware of a careful analysis of this question.

and

$$\hat{u}_{it} = b_{it} + \frac{\widehat{G}_{M,B}(b_{it}, b_{it}; \mathcal{N}_t)}{\hat{g}_{M,B}(b_{it}, b_{it}; \mathcal{N}_t)}$$

to obtain consistent estimators under standard conditions. In practice, however, this approach may require a great deal of data, since many observations will be needed for each set $\mathcal{N}$ considered.

In some cases, one may be able to categorize bidders into a smaller set of heterogeneous classes, assuming exchangeability within each class. This structure can lead to significant practical advantages, as it allows use of substantially more data for each estimated pseudo-value. For example, Campo, Perrigne and Vuong (2003) studied "wildcat" auctions for mineral extraction rights on the US outer-continental shelf, where bids may come from "solo" bidders (a single firm) or "joint" bidders (more than one firm, legally bidding as one).[23] This leads them to consider the case of two classes of bidders, I and II. The first-order condition for a class-I bidder in an auction in which the set of bidders is $\mathcal{N}$ can be written

$$u_{\mathrm{I}} = b_{\mathrm{I}} + \frac{G^{\mathrm{I}}_{M,B}(b_{\mathrm{I}}, b_{\mathrm{I}}; \mathcal{N})}{g^{\mathrm{I}}_{M,B}(b_{\mathrm{I}}, b_{\mathrm{I}}; \mathcal{N})}. \tag{3.6}$$

Define the relation $=^{\mathrm{I,II}}$ such that $\mathcal{N}_t =^{\mathrm{I,II}} \mathcal{N}$ holds iff $\mathcal{N}_t$ and $\mathcal{N}$ have the same number of bidders, n^{I} and n^{II}, from each class. Let $T^{\mathrm{I,II}}_{\mathcal{N}} = \sum_{t=1}^{T} \mathbf{1}\{\mathcal{N}_t =^{\mathrm{I,II}} \mathcal{N}\}$. Now $G^{\mathrm{I}}_{M,B}(b_{\mathrm{I}}, b_{\mathrm{I}}; \mathcal{N})$ can be estimated by

$$\widehat{G}^{\mathrm{I}}_{M,B}(b, b; \mathcal{N}) = \frac{1}{T^{\mathrm{I,II}}_{\mathcal{N}} \times h_G \times n^{\mathrm{I}}} \sum_{s=1}^{T} \sum_{i=1}^{|\mathcal{N}|} K\left(\frac{b - b_{is}}{h_G}\right) \mathbf{1}\{m_{is} < b,\ \mathcal{N}_s =^{\mathrm{I,II}} \mathcal{N},\ i \in \text{class I}\}.$$

Analogous adjustments are made to an estimator for $g^{\mathrm{I}}_{M,B}(b, b; \mathcal{N})$ and to the first-order condition for a class-II bidder [see Campo, Perrigne and Vuong (2003, pp. 186–187) for details]. Note in particular that in estimating $\frac{G^{\mathrm{I}}_{M,B}(b_{\mathrm{I}}, b_{\mathrm{I}}; \mathcal{N})}{g^{\mathrm{I}}_{M,B}(b_{\mathrm{I}}, b_{\mathrm{I}}; \mathcal{N})}$ one can use data from all auctions t with $\mathcal{N}_t =^{\mathrm{I,II}} \mathcal{N}$. Furthermore, the sample of bids is cut less finely across bidders than in the completely general case.

Note that we have treated asymmetries as resulting from differences in the distributions from which bidders draw unobservables. In some cases, it may be more natural that asymmetries arise instead from observable covariates Z_i that are idiosyncratic to each bidder – e.g., distance to a construction site [e.g., Bajari (1997), Flambard and

[23] Athey, Levin and Seira (2004) provide another example, treating loggers and sawmills as two different classes of bidders at timber auctions.

Perrigne (2006)]. Conditional on having the same value of the covariates, bidder valuations may still be exchangeable. Without further restriction, this is similar to the case in which bidders fall into discrete categories; indeed, it is exactly the same if the covariates are discrete. In the case of continuous covariates, standard smoothing techniques would lead to similar approaches for estimating the joint distribution $F_{\mathbf{X},\mathbf{U}}(\cdot|Z_1,\ldots,Z_n)$. In Section 6.2.1 we will see how the presence of bidder-specific covariates can actually aid identification in some cases.

3.3. Incomplete bid data and Dutch auctions

3.3.1. Independent private values

The results above exploited the assumed observability of all bids from each auction. In some applications, however, not all bids are available. For example, for some auctions only the transaction price $B^{(n:n)}$ is recorded. One example is a Dutch auction, where the auctioneer starts with a very high price and lowers it continuously until one bidder is willing to take the good at the current price. Although a Dutch auction is seemingly different from a first-price sealed-bid auction, the two formats are strategically equivalent (assuming the same information is observable prior to bidding).[24] Since a Dutch auction ends as soon as the winner makes his bid, only the winning bid can be observed. We will see that in some cases the winning bid is sufficient for identification. In other environments, only a partial set of bids may be available. For example, in a procurement setting, the buyer might retain information regarding the best losing bid in case the auction winner defaults. Viewed somewhat differently, identification results for the case of incomplete bid data can clarify how much information one would need to collect to create a useful data set.

In an asymmetric IPV first-price (or Dutch) auction, identification of each marginal distribution $G_{B_i}(\cdot)$ from observation of the winning bid and winner's identity is formally equivalent to identification of the well known competing risks model with independent nonidentically distributed risks.[25] For that model, nonparametric identification was shown by Berman (1963). Since knowledge of each $G_{B_i}(\cdot)$ completely determines the distribution of

$$B_i + \frac{G_{M_i|B_i}(B_i|B_i;\mathcal{N})}{g_{M_i|B_i}(B_i|B_i;\mathcal{N})}$$

[24] Brendstrup and Paarsch (2003) point out that in a Dutch auction the set of actual opponents may be observable before the bidding decision is made. With a binding reserve price that creates a distinction between the potential bidders and actual bidders (see Section 6.3 for definitions), this destroys strategic equivalence. The basic approach for first-price auctions can still be applied, however, if the distribution of the actual bidders' valuations (which reflects truncation at the reserve price) is the object of interest. See the related discussion in Section 6.3.1.

[25] The data generating process mapping bids to observables is formally identical to that in a complementary risks model, where failure of all components triggers the observable system failure, and one observes the identity of the last component to fail. This is isomorphic to the competing risks model.

identification of the marginal distributions $F_{U_i}(\cdot)$ then follows. This gives the following result from Athey and Haile (2002).

THEOREM 3.2. *Suppose that the transaction price and the number of bidders* (*and, if bidders are asymmetric, the set* $\mathcal{N}$ *and identity of the winner*) *are observed in first-price auctions with independent private values. Then* $F_{\mathbf{U}}(\cdot)$ *is identified.*

To gain some intuition, consider the symmetric case, where the observable transaction price $B^{(n:n)}$ has distribution $G_B^{(n:n)}(b)$, $G_{B_i}(\cdot)$ can be written as $G_B(\cdot)$, and

$$\frac{G_B^{(n:n)}(b)}{g_B^{(n:n)}(b)} = \frac{G_B(b)^n}{n g_B(b) G_B(b)^{n-1}} = \frac{n-1}{n}\left(\frac{G_B(b)}{(n-1)g_B(b)}\right). \tag{3.7}$$

As first observed by Guerre, Perrigne and Vuong (1995), identification then follows from (3.5).

In the asymmetric case, the derivation of Berman's (1963) equation (2) [see also Prakasa-Rao (1992, Theorem 7.3.1 and Remarks 7.3.1)] yields the relation (fixing $\mathcal{N}$)

$$G_{B_i}(b_i) = \exp\left\{\int_{-\infty}^{b_i}\left(\sum_{j=1}^{n} G_i^w(s)\right)^{-1} \mathrm{d}G_j^w(s)\right\}, \tag{3.8}$$

where $G_i^w(b_i) = \Pr(B_i \leqslant b_i, B_i \geqslant B_j\ \forall j)$. Since each $G_i^w(b_i)$ is observable, each $G_{B_i}(b_i)$ is identified. The marginal distributions $G_{B_i}(\cdot)$ uniquely determine the underlying distributions $F_{U_i}(\cdot)$ through the first-order condition (3.1) as in the case in which all bids are observed.

An immediate implication of Theorem 3.2 is identification from the transaction price in a Dutch auction.

COROLLARY 3.1. *Suppose that the transaction price and the number of bidders* (*and, if bidders are asymmetric, the set* $\mathcal{N}$ *and the identity of the winner*) *are observed in Dutch auctions with independent private values. Then* $F_{\mathbf{U}}(\cdot)$ *is identified.*

As suggested by Laffont, Ossard and Vuong (1995), the requirement that n be observable by the econometrician may fail in some Dutch auctions, where one might expect only the transaction price (i.e., the only bid made in the auction) to be recorded. It should be clear that without knowledge of n, knowledge of $G_B^{(n:n)}(\cdot)$ is insufficient to determine even $G_B(\cdot)$.[26]

Extending the estimation approaches described in the preceding sections to cases in which only the transaction price (winning bid) is observed is straightforward in the symmetric case, where (3.7) can be used. In the asymmetric case, Brendstrup and Paarsch

[26] Laffont, Ossard and Vuong (1995) suggest an approach for estimating n when it is unknown but fixed. They assume that identification follows from a parametric distributional assumption.

(2003) have recently proposed substituting the empirical distribution function and a kernel density estimator for, respectively, $G_i^w(s)$ and $\frac{\mathrm{d}G_i^w(s)}{\mathrm{d}s}$ in Equation (3.8). The close relation of the model to the competing risks model suggests that other nonparametric estimators such as the Nelson–Aalan or Kaplan–Meier estimators [e.g., David and Moeschberger (1978), Andersen et al. (1991)] might also be used to estimate each $G_{B_i}(\cdot)$. Once these distributions are estimated, one might then simulate bids from these estimated distributions in order to estimate pseudo-values using the first-order condition (3.1).

3.3.2. Affiliated private values

Next we consider what can be learned from the top two bids in first-price auctions in a richer private values environment. Intuitively, the top two bids contain much of the critical information for a first-price auction. First, these are the only two bids necessary to determine the distribution of the maximum opposing bid for each bidder, suggesting that at least some information about the markdown components of the equilibrium bid functions could be learned. Second, in equilibrium, the top two bids are monotonic transformations of the top two signals. As the following result, adapted from Athey and Haile (2002) shows, this is sufficient to enable partial identification in a symmetric first-price sealed-bid auction.

THEOREM 3.3. *Assume that the two highest bids are observed in first-price auctions. If bidders are asymmetric, assume that the set $\mathcal{N}$ and the identity of the winner are also observed. Then*

(i) *the equilibrium bid functions $\beta_i(\cdot;\mathcal{N})$ are identified for all $i = 1,\dots,n$;*
(ii) *with symmetric private values, the joint distribution of $U^{(n-1:n)}$ and $U^{(n:n)}$ is identified.*

PROOF. Part (ii) follows immediately from part (i), since in the symmetric private values case the two highest bids are made by the bidders with the two highest valuations. To prove part (i) for the more general asymmetric case, consider bidder 1 without loss of generality. Let $I^{(n:n)}$ denote the identity of the winning bidder. For almost all such $b_1 \in \operatorname{supp}[G_{B_1}(\cdot)]$ (using Bayes' rule, and canceling common terms)

$$
\begin{aligned}
&\frac{\Pr(\max_{j\neq 1} B_j \leqslant b_1 \mid B_1 = b_1;\mathcal{N})}{\frac{\partial}{\partial m}\Pr(\max_{j\neq 1} B_j \leqslant m \mid B_1 = b_1;\mathcal{N})|_{m=b_1}} \\
&\quad= \frac{\frac{\partial}{\partial y}\Pr(\max_{j\neq 1} B_j \leqslant b_1,\ B_1 \leqslant y;\mathcal{N})|_{y=b_1}}{\frac{\partial^2}{\partial m\partial y}\Pr(\max_{j\neq 1} B_j \leqslant m,\ B_1 \leqslant y;\mathcal{N})|_{m=y=b_1}} \\
&\quad= \frac{\frac{\partial}{\partial y}G_{\mathbf{B}}(y,b_1,\dots,b_1;\mathcal{N})|_{y=b_1}}{\sum_{j\neq 1}\frac{\partial^2}{\partial y\partial s_j}G_{\mathbf{B}}(y,s_2,\dots,s_n;\mathcal{N})|_{y=s_2=\cdots=s_n=b_1}}
\end{aligned}
$$

$$= \frac{\frac{\partial}{\partial y}\Pr(B^{(n:n)} \leqslant y,\ I^{(n:n)} = 1, \mathcal{N})|_{y=b_1}}{\frac{\partial^2}{\partial m \partial y}\Pr(B^{(n-1:n)} \leqslant m,\ B^{(n:n)} \leqslant y,\ I^{(n:n)} = 1, \mathcal{N})|_{m=y=b_1}}.$$

Since the last expression is the ratio of two observable functions, the right-hand side of (2.4) is identified almost everywhere, which determines bidder 1's (inverse) equilibrium bid function. □

Estimation approaches based on this partial identification result have not yet been explored. Note that estimates of each $\beta_i(\cdot; \mathcal{N})$ are themselves of interest, since these characterize the wedge between bids and valuations that determine the division of surplus and can lead to inefficiencies. In the symmetric case, knowledge of $\beta(\cdot; n)$ and the joint distribution of $(U^{(n-1:n)}, U^{(n:n)})$ would enable evaluation of rent extraction by the seller, the effects of introducing a reserve price, and the outcomes under a number of alternative selling mechanisms. As discussed in Section 8, this partial identification result can also be sufficient to enable discrimination between private and common values environments.

Observing the top two bids, however, is insufficient to identify the full joint distribution $F_{\mathbf{U}}(\cdot)$. In fact, Athey and Haile (2002) have shown that observation of *all* bids is needed, even in a symmetric setting.

THEOREM 3.4. *In the symmetric private values model, suppose that* $(B^{(n:n)}, B^{(n-1:n)})$ *are observed in first-price auctions but some* $B^{(j:n)}$, $j < n-1$ *is unobserved. Then* $F_{\mathbf{U}}(\cdot)$ *is not identified.*

PROOF. Let the point $(u_1, u_2, \ldots, u_n)$ be on the interior of the support of $F_{\mathbf{U}}(\cdot)$, with $u_1 < \cdots < u_n$. Starting with the true joint density $f_{\mathbf{U}}(\cdot)$, define a new joint density $\tilde{f}_{\mathbf{U}}(\cdot)$ by shifting mass δ from a neighborhood of $(u_1, \ldots, u_j, \ldots, u_n)$ (and each permutation) to a neighborhood of the point $(u_1, \ldots, u_j + \epsilon, \ldots, u_n)$ (and each permutation).[27] For small ϵ and δ, this change preserves exchangeability. Since the distribution of $\max_{k \neq i} B_k$ is unaffected for any i by this change, equilibrium bidding strategies (given by (3.1)) remain the same for all bidders. Furthermore, the only order statistic affected in moving from $\tilde{f}_{\mathbf{U}}(\cdot)$ to $f_{\mathbf{U}}(\cdot)$ is $U^{(j:n)}$. Since $B^{(j:n)} = \beta(U^{(j:n)}; n)$ is unobserved, the distribution of observables is unchanged. □

Intuitively, even under exchangeability, $F_{\mathbf{U}}(\cdot)$ is an n-dimensional joint distribution. Identifying this distribution with data of dimension $n-1$ or lower would require additional restrictions.

[27] Athey and Haile (2002, Theorem 4) describe this in more detail.

4. Ascending auctions with private values: Basic results

4.1. Identification

With private values, equilibrium bidding strategies in Milgrom and Weber's (1982) model of the ascending auction are particularly simple: it is a weakly dominant strategy for each bidder to exit the auction at his valuation. Hence, unlike the first-price auction, here there is no need to estimate inverse bid functions in order to relate the observed bids to the underlying valuations. This does not make identification trivial, however. The reason is the fact that the auction ends as soon as only one bidder remains. Because the auction stops at the second highest bid, the only information revealed about the winner's valuation is the censoring point $B^{(n-1:n)}$. This partial observability of bids is the main challenge to identification.

When valuations are independent, Athey and Haile (2002) have shown that identification does hold, even if one observes only the transaction price (and the identity of the winner, if bidders are asymmetric). This is easier to see when bidders are symmetric. In that case valuations are i.i.d. draws from the marginal distribution $F_U(\cdot)$. The transaction price is the order statistic $U^{(n-1:n)}$, which has distribution $F_U^{(n-1:n)}(\cdot)$. The distribution of an order statistic from an i.i.d. sample of size n from an arbitrary distribution $F(\cdot)$ has the distribution [see, e.g., Arnold, Balakrishnan and Nagaraja (1992)]

$$F^{(i:n)}(s) = \frac{n!}{(n-i)!(i-1)!} \int_0^{F(s)} t^{i-1}(1-t)^{n-i}\,\mathrm{d}t \quad \forall s. \tag{4.1}$$

Since the right-hand side of (4.1) is strictly increasing in $F(s) \in [0,1]$, $F^{(i:n)}(s)$ uniquely determines $F(s)$ for every s.

When bidders have asymmetric independent private values, the identification argument is more subtle. Athey and Haile (2002) point out that the asymmetric ascending auction model is isomorphic to a model considered in the statistics literature on reliability, where Meilijson (1981) has provided a proof. To get some intuition, fix $\mathcal{N}$ with $|\mathcal{N}| = 3$ and define

$$\begin{aligned}
\widetilde{G}_3(t) &\equiv \Pr(\text{price} \leqslant t,\ 3 \text{ is the winner}) \\
&= \Pr(B_1 \leqslant t;\ B_2 \leqslant t;\ B_3 \geqslant t) + \Pr(B_1 \leqslant B_3;\ B_2 \leqslant B_3;\ B_3 \leqslant t) \\
&= F_{U_1}(t)F_{U_2}(t)\big(1 - F_{U_3}(t)\big) + \int_{\underline{u}}^{t} F_{U_1}(x)F_{U_2}(x)\,\mathrm{d}F_{U_3}(x) \\
&= \int_{\underline{u}}^{t} \big(F_{U_1}(x)F_{U_2}(x)\big)'\big(1 - F_{U_3}(x)\big)\,\mathrm{d}x,
\end{aligned}$$

where $(F_{U_1}F_{U_2})'$ is the first derivative of $F_{U_1}F_{U_2}$ and the last equality follows from integration by parts. Differentiating both sides, we obtain

$$\tilde{g}_3(t) = \big(F_{U_1}(t)F_{U_2}(t)\big)'\big(1 - F_{U_3}(t)\big),$$

which implies that

$$\left(F_{U_1}(t)F_{U_2}(t)\right)' = \frac{\tilde{g}_3(t)}{1-F_{U_3}(t)},$$

$$F_{U_1}(t)F_{U_2}(t) = \int_{\underline{u}}^{t} \frac{\tilde{g}_3(x)}{1-F_{U_3}(x)}\,\mathrm{d}x,$$

$$\log F_{U_1}(t) + \log F_{U_2}(t) = \log \int_{\underline{u}}^{t} \frac{\tilde{g}_3(x)}{1-F_{U_3}(x)}\,\mathrm{d}x.$$

Rewrite this as

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \log F_{U_1}(x) \\ \log F_{U_2}(x) \\ \log F_{U_3}(x) \end{bmatrix} = \log \begin{bmatrix} \int_{\underline{u}}^{t} \frac{\tilde{g}_3(x)}{1-F_{U_3}(x)}\,\mathrm{d}x \\ \int_{\underline{u}}^{t} \frac{\tilde{g}_2(x)}{1-F_{U_2}(x)}\,\mathrm{d}x \\ \int_{\underline{u}}^{t} \frac{\tilde{g}_1(x)}{1-F_{U_1}(x)}\,\mathrm{d}x \end{bmatrix}. \tag{4.2}$$

This is a 3×3 system of operator equations defining how the three observable marginal distributions are related to the three marginal distributions $F_{U_i}(\cdot)$ of interest. Meilijson (1981) showed that this system has a unique solution.

We summarize these results in the following theorem.

THEOREM 4.1. *In an ascending auction with symmetric independent private values, $F_{\mathbf{U}}(\cdot)$ is identified when the transaction price and the number of bidders are observable. In an ascending auction with asymmetric independent private values, $F_{\mathbf{U}}(\cdot)$ is identified when the transaction price, the identity of the winning bidder, and the set $\mathcal{N}$ are observable.*

One attractive feature of this result is that it implies that one need not use bids other than the transaction price to estimate $F_U(\cdot)$. This is valuable because in many applications one may have little confidence in the interpretation of losing bids implied by Milgrom and Weber's (1982) button auction model. With independence, one is free to ignore losing bids altogether, relying only on the assumption that the transaction price equals the second highest valuation.

Athey and Haile (2002) give a much more negative result when the independence assumption is dropped, even with symmetric bidders. The proof mirrors that of Theorem 3.4.

THEOREM 4.2. *In a symmetric private values model, the joint distribution $F_{\mathbf{U}}(\cdot)$ is not identified from the observable bids in an ascending auction.*

4.2. Estimation

Here we make the same sampling assumptions made in the discussion of estimation for first-price sealed bid auctions (see Section 3.2). In an ascending auction, typically

one treats the highest price offered by each bidder as his "bid," i.e., his exit price in the button auction model.[28] Using these data, several parametric estimation approaches for the symmetric IPV model have been explored in the literature. Maximum likelihood, nonlinear least squares, and GMM are among the methods considered. Due to the simplicity of the equilibrium bid function, likelihood functions or moment conditions are easily derived from the probability density function of the winning bid alone [e.g., Paarsch (1992b), Donald and Paarsch (1996), Baldwin, Marshall and Richard (1997), Haile (2001)], or of the $n-1$ losing bids [e.g., Donald and Paarsch (1996), Paarsch (1997)]. In the former case, the likelihood of a winning bid b is $f_U^{(n-1:n)}(b)$. For the latter case, the likelihood for the losing bids in a given auction is

$$n!\big[1 - F_U\big(b^{(n-1:n-1)}\big)\big] \prod_{j<n} f_U\big(b^{(j:n)}\big).$$

To our knowledge, nonparametric estimation of the symmetric IPV model has been performed only in simulations [Haile and Tamer (2003)], although this is actually simpler than nonparametric estimation in the case of a first-price auction. Following Haile and Tamer (2003), for $H \in [0,1]$ define the strictly increasing differentiable function $\phi(H;i,n)$ implicitly as the solution to

$$H = \frac{n!}{(n-i)!(i-1)!} \int_0^{\phi} s^{i-1}(1-s)^{n-i}\,\mathrm{d}s \tag{4.3}$$

so that by (4.1)

$$F_U(u) = \phi\big(F_U^{(i:n)}(u); i, n\big) \quad \forall u, i \leqslant n. \tag{4.4}$$

In particular,

$$F_U(u) = \phi\big(F_U^{(n-1:n)}(u); n-1, n\big) \quad \forall u. \tag{4.5}$$

Since the winning bid is $B^{(n-1:n)} = U^{(n-1:n)}$, one can construct an estimator of $F_U(u)$ by substituting the empirical distribution

$$\widehat{G}_B^{(n-1:n)}(u) = \frac{1}{T_n} \sum_{t=1}^{T} \mathbf{1}\big\{n_t = n,\ B^{(n_t-1:n_t)} \leqslant u\big\}$$

for $F_U^{(n-1:n)}(u)$ inside the function $\phi(\cdot)$ on the right-hand side of (4.5). Since $G_B^{(n-1:n)}(\cdot) = F_U^{(n-1:n)}(\cdot)$, by standard arguments $\widehat{G}_B^{(n-1:n)}(u)$ converges uniformly to $F_U^{(n-1:n)}(u)$ almost surely and has a normal asymptotic approximation. Convergence of

[28] See, e.g., the surveys of Paarsch (1994) and Hendricks and Paarsch (1995). A source of ambiguity arises when one observes n such bids, with $B^{(n:n)}$ significantly higher than $B^{(n-1:n)}$. In such cases, one may question the applicability of the button auction model. For now we assume the button auction structure and treat the distributions $G_B^{(n:n)}(\cdot)$ and $G_B^{(n-1:n)}(\cdot)$ as identical.

$\phi(\widehat{F}_U^{(n-1:n)}(u); n-1, n)$ to $F_U(u)$ then follows, with an asymptotic normal distribution obtainable by the delta method.[29] In practice, one can use the relation

$$F_U^{(n-1:n)}(u) = \sum_{j=n-1}^{n} \binom{n}{j} F_U(u)^j \big(1 - F_U(u)\big)^{n-j} \tag{4.6}$$

instead of the equivalent but more computationally demanding (4.5) when solving for each $\widehat{F}_U(u)$. Monotonicity of the relation between $F_U^{(n-1:n)}(u)$ and $F_U(u)$ makes numerical solution particularly simple.

Note that when there is exogenous variation in the number of bidders, there will be as many different estimators $\phi(\widehat{F}_U^{(n-1:n)}(u); n-1, n)$ of $F_U(u)$ available as there are different values of n in the data. If one observes losing bids beyond the transaction price and assumes these are generated by the button auction model, additional estimators will be available, based on Equation (4.4) with $i < n-1$. An efficient estimator would take an optimally weighted average of these different estimators, imposing the constraint that the estimated CDF be monotone.

In the case of asymmetric bidders, no simple relation like (4.6) is available. However, a likelihood approach provides several possible estimation strategies. The likelihood for the observable event $\{i$ wins at price $p\}$ is

$$\big(1 - F_{U_i}(p)\big) \sum_{j \neq i} f_{U_j}(p) \prod_{k \neq i,j} F_{U_k}(p).$$

Hence, if we let I_t denote the winner of auction t, the likelihood function has the form

$$\mathcal{L} = \prod_t \big(1 - F_{U_{I_t}}(p)\big) \sum_{j \neq I_t} f_{U_j}(p) \prod_{k \neq I_t, j} F_{U_k}(p). \tag{4.7}$$

Parametric or nonparametric MLE might then be applied. Brendstrup and Paarsch (2004) have recently proposed a "semi-nonparametric" [Gallant and Nychka (1987)] estimation approach based on this likelihood.[30]

4.3. An alternative, incomplete model of ascending auctions

In some cases an auction institution closely matching the structure of the button auction model is observed in practice. Bidders may, for example, raise their hands or other objects to indicate their participation continuously as the auctioneer raises the price [see, e.g., Zulehner's (2003) description of cattle auctions]. When the auction is conducted in a less structured oral format, however, one may question the applicability of the button

[29] In fact, the convergence is uniform. These results follow from those given in Haile and Tamer (2002, Appendix A; 2003, Theorem 3).

[30] They also consider auctions in which multiple units are sold sequentially, focusing on bids in the (single-unit, asymmetric) auction of the final unit.

auction model as an empirical structure. Of particular concern is the fact that there is no need for a bidder to continuously indicate whether she is "in" or "out" as the auction proceeds. Nontrivial minimum bid increments are often used, and a bidder is free to "jump bid" or to remain silent for most of the auction and bid only when it looks like the auction is about to end (if others do not bid the price past her valuation first). Such behaviors are common in practice and raise the possibility that bidders will fail to reveal their full willingness to pay, or even fail to bid altogether. Several theoretical extensions of the standard model have been proposed, mainly focusing on the case of common values [Avery (1998), Harstad and Rothkopf (2000), Izmalkov (2003)]. Until recently, however, all empirical models of the ascending auction relied on significant abstractions for tractability of the underlying theoretical model.

As an alternative to relying on the structure of the button auction or another stylized model, Haile and Tamer (2003) have proposed an empirical approach to ascending auctions with symmetric independent private values using two simple assumptions to govern the interpretation of the observed bids:

ASSUMPTION 4.1. Bidders do not bid more than they are willing to pay.

ASSUMPTION 4.2. Bidders do not allow an opponent to win at a price they are willing to beat.

These assumptions allow bidding as in the dominant strategy equilibrium of the button auction model but do not require it. In particular, bids need not be equal to valuations or even monotonic in valuations, and the price need not equal the second highest valuation. These assumptions define an "incomplete" model of an ascending auction: they place some restrictions on the relation between valuations and bids, but do not fully characterize behavior. While this incomplete model is insufficient to identify the distribution of valuations from the distribution of bids, it does provide partial identification; in particular, one may still obtain informative *bounds* on the distribution of valuations.

4.3.1. Bounding the distribution of bidder valuations

To obtain an upper bound on the distribution function $F_U(\cdot)$, observe that Assumption 4.1 is equivalent to assuming $b_i \leqslant u_i$ for all i. In an n-bidder auction, it is easy to confirm that this implies $b^{(i:n)} \leqslant u^{(i:n)}\ \forall i$, which then gives

$$G_B^{(i:n)}(u) \geqslant F_U^{(i:n)}(u) \quad \forall i, n, u. \tag{4.8}$$

Recalling the definition (4.3) and Equation (4.4), we know that

$$F_U(u) = \phi\big(F_U^{(i:n)}(u); i, n\big) \quad \forall i, n, u. \tag{4.9}$$

Since the function $\phi(\cdot; i, n) : [0, 1] \to [0, 1]$ is strictly increasing, (4.8) and (4.9) together give

$$\phi\big(G_B^{(i:n)}(u); i, n\big) \geqslant F_U(u) \quad \forall i, n, u.$$

For each u, this yields $\sum_{n=\underline{n}}^{\bar{n}} n$ upper bounds on $F_U(u)$. The most informative bound (i.e., the smallest upper bound) is obtained by taking the minimum at each value of u:

$$F_U^+(u) = \min_{i,n} \phi\big(G_B^{(i:n)}(u); i, n\big). \tag{4.10}$$

A similar approach can be taken to obtain a lower bound on $F_U(\cdot)$. Letting $\Delta \geqslant 0$ denote the minimum bid increment in the auction, Assumption 4.2 implies that all losing bidders have valuations less than $b^{(n:n)} + \Delta$, implying $u^{(n-1:n)} \leqslant b^{(n:n)} + \Delta$. If $G_\Delta^{(n:n)}(\cdot)$ denotes the distribution of $B^{(n:n)} + \Delta$, this gives

$$G_\Delta^{(n:n)}(u) \leqslant F_U^{(n-1:n)}(u) \quad \forall n, u.$$

Applying the monotonic transformation $\phi(\cdot; n-1, n)$ to both sides gives

$$\phi\big(G_\Delta^{(n:n)}(u); n-1, n\big) \leqslant F_U(u) \quad \forall n, u.$$

This yields multiple lower bounds on $F_U(u)$ (one for each value of n). The most informative bound can be constructed by taking the pointwise maximum:

$$F_U^-(u) = \max_n \phi\big(G_\Delta^{(n:n)}(u); n-1, n\big). \tag{4.11}$$

We summarize these results in the following theorem.

THEOREM 4.3. *$F_U^-(u) \leqslant F_U(u) \leqslant F_U^+(u)$ for all u.*

Note that in principle this approach can be followed even when the transaction price is the only bid available from each auction – the only modification required is that the minimum in (4.10) would be taken over n only, fixing $i = n$. However, an essential requirement of the approach is that the number of bidders, n, be observable to the econometrician. This is also essential for the methods discussed above for both sealed-bid and button auction models, but the assumption may be more suspect in an ascending auction in which the button auction structure is inappropriate. The number n will be observed if all bidders make some bid during the auction, or if bidders must pre-qualify, register, or otherwise identify themselves in order to be eligible to bid. This is the case in the timber auctions studied by Haile and Tamer (2003) and many other public sector auctions.[31]

In general, the informativeness of the bounds $F_U^+(u)$ and $F_U^-(u)$ depends on the deviation of the true data generating process from that implied by the button auction model. In fact, if the restriction $B^{(n:n)} = B^{(n-1:n)}$ implied by the button auction model is consistent with the data, the bounds $F_U^-(\cdot)$ and $F_U^+(\cdot)$ collapse to the true distribution $F_U(\cdot)$, providing point identification. By contrast, imposing the full structure of the button auction model when this is not the true data generating process need not result in

[31] Song (2004) has recently considered identification and estimation for ascending auctions (and others) when n is not observed. We will discuss one such case in Section 6.3.4 below.

an estimate of $F_U(\cdot)$ that lies within the bounds, regardless of sample size.[32] While this should not be surprising – imposing false restrictions should be expected to yield misleading estimates – it is a useful reminder that imposing structure in order to obtain point identification is not equivalent to selecting a point estimate within bounds obtained from a less restrictive but incomplete model.

Estimation of the bounds is achieved by substituting the empirical distributions

$$\widehat{G}_B^{(i:n)}(b) = \frac{1}{T_n}\sum_{t=1}^{T} \mathbf{1}\{n_t = n,\ b^{(i:n_t)} \leqslant b\}$$

and

$$\widehat{G}_\Delta^{(n:n)}(b) = \frac{1}{T_n}\sum_{t=1}^{T} \mathbf{1}\{n_t = n,\ b^{(n_t:n_t)} + \Delta_t \leqslant b\}$$

for the corresponding CDFs in (4.10) and (4.11). Since the empirical distribution functions are uniformly consistent and asymptotically normally distributed estimators of their population analogs, differentiability of $\phi(\cdot; i, n)$ ensures that each $\phi(G_B^{(i:n)}(u); i, n)$ and $\phi(G_\Delta^{(n:n)}(u); n-1, n)$ are consistent and asymptotically normally distributed as well. Continuity of the min and max functions then ensures consistency of the estimates of the estimators

$$\widehat{F}_U^+(u) = \min_{i,n} \phi\big(\widehat{G}_B^{(i:n)}(u); i, n\big),$$

$$\widehat{F}_U^-(u) = \max_{n} \phi\big(\widehat{G}_\Delta^{(n:n)}(u); n-1, n\big).$$

These estimators have nonnormal asymptotic distributions, due to the max and min. However, Haile and Tamer (2002) show that the bootstrap (see Section 3.2.1) may be used for inference. A more difficult problem is that, while these estimators are consistent, in practice the max and min can lead to severe finite sample bias, potentially even leading to estimated upper and lower bounds that cross. Intuitively, taking the minimum (maximum) of several consistent estimators makes it likely that an estimator with downward (upward) sampling error is selected. One solution, discussed in greater detail by Haile and Tamer (2003), is to define bounds in finite samples based on smooth approximations to the max and min functions in the definitions of $\widehat{F}_U^+(u)$ and $\widehat{F}_U^-(u)$ above. This amounts to using weighted averages instead of the max or min.

4.3.2. *Bounding the optimal reserve price*

Unlike point estimates of $F_U(\cdot)$, it is not immediately clear whether bounds on $F_U(\cdot)$ would be useful.[33] For example the key policy choice for the seller in the symmetric

[32] See Haile and Tamer (2003) for additional discussion and simulation results.

[33] Haile and Tamer (2003) demonstrate an additional use of the bounds by showing how to incorporate auction covariates nonparametrically. Building on Manski and Tamer (2002), the resulting bounds on conditional

IPV environment is the reserve price [Myerson (1981)]. When $F_U(\cdot)$ is continuously differentiable, the optimal reserve price r^* is defined by the equation[34]

$$r^* = c_0 + \frac{1 - F_U(r^*)}{f_U(r^*)}, \tag{4.12}$$

where c_0 is the seller's valuation (or marginal cost) of the good. However, nondegenerate bounds on $F_U(\cdot)$ place no restriction on its derivative $f_U(\cdot)$ at any given point. Hence, just as a monopolist's price need not shift in the same direction as demand, r^* need not lie between the reserve prices that would be optimal if $F_U^+(\cdot)$ or $F_U^-(\cdot)$ were the true distribution of valuations. Note that the same problem arises any time one wishes to construct confidence bands on the optimal reserve price from confidence bands on nonparametric point estimates of $F_U(u)$, e.g., using the method described in Section 4.2.

Observe, however, that when the seller's own valuation for the good is c_0, r^* solves $\max_r \pi(r)$, where[35]

$$\pi(r) = (r - c_0)\big(1 - F_U(r)\big).$$

Since $F_U(r)$ must lie between $F^-(r)$ and $F^+(r)$, $\pi(r)$ must lie between

$$\pi_1(r) = (r - c_0)\big(1 - F_U^+(r)\big)$$

and

$$\pi_2(r) = (r - c_0)\big(1 - F_U^-(r)\big).$$

Figure 1 illustrates. Under the additional assumption that $\pi(r)$ is strictly quasi-concave in r (which ensures a unique solution to (4.12)) we can use the bounding "profit" functions $\pi_1(\cdot)$ and $\pi_2(\cdot)$ to place bounds on r^*. Let $r_1^* \in \arg\sup \pi_1(r)$, $r_2^* \in \arg\sup \pi_2(r)$, and $\pi_1^* = \pi_1(r_1^*)$. We obtain the trivial result $r^* = r_1^*$ when $\pi_2(r_1^*) = \pi_2(r_2^*) = \pi_1^*$, or when $\pi_2(r_1^*) = \pi_1^*$ and either $\pi_1(\cdot)$ or $\pi_2(\cdot)$ has slope zero at r_1^*. For these trivial cases let $r^- = r^+ = r_1^*$. For all other cases define

$$r^- = \sup\big\{r < r_1^*\colon\ \pi_2(r) \leqslant \pi_1^*\big\},$$
$$r^+ = \inf\big\{r > r_1^*\colon\ \pi_2(r) \leqslant \pi_1^*\big\}.$$

Haile and Tamer (2003) prove the following result.[36]

distributions can then be used to estimate bounds on parameters of a semiparametric model describing how valuations shift with auction characteristics.

[34] This is easily derived for a second-price sealed-bid or button auction, where a reserve price of r implies expected revenue $rnF_U(r)^{n-1}(1 - F_U(r)) + \int_t^\infty un(n-1) f_U(u) F_U(u)(1 - F_U(u))\,\mathrm{d}u$. Myerson (1981) shows that, under a regularity condition, a standard auction with an optimal reserve price is optimal among all possible selling mechanisms. Haile and Tamer (2003) show that r^* is also optimal in their incomplete model ascending auction as long as Assumptions 4.1 and 4.2 are interpreted as a partial characterization of equilibrium behavior in some true but unspecified auction mechanism.

[35] Note that $\pi(r)$ is not the expected profit of the seller when $n > 1$. The usefulness of this function is the fact that its maximum is attained at the same value of r that maximizes the seller's expected profit.

[36] Bounds are said to be *sharp* if they exhaust all information available from the data and *a priori* assumptions.

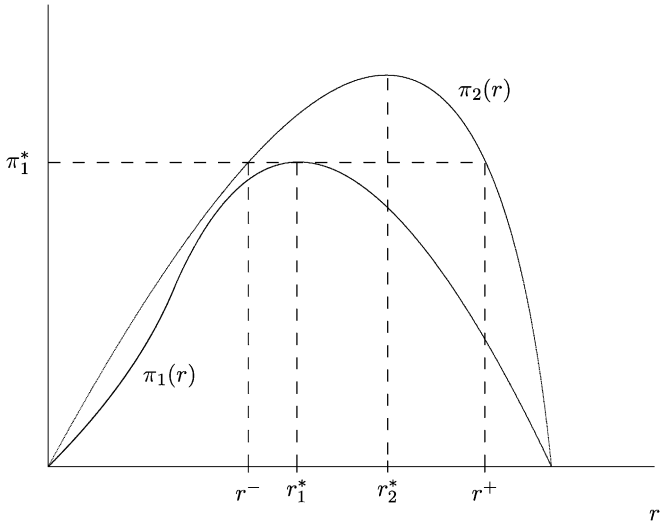

Figure 1.

THEOREM 4.4. *Suppose $\pi(r)$ is continuously differentiable and strictly quasi-concave in r. Then $r^* \in [r^-, r^+]$. Given the bounds $F^+(\cdot)$ and $F^-(\cdot)$ on $F_U(\cdot)$, the bounds r^- and r^+ on r are sharp.*

Intuition for the result can be seen in Figure 1. We know that the true function $\pi(\cdot)$ lies between $\pi_1(\cdot)$ and $\pi_2(\cdot)$ and must, therefore, reach a peak of at least π_1^*. Such a peak cannot be reached outside the interval $[r^-, r^+]$. However, prices arbitrarily close to either of these endpoints could be the true optimum r^*.

For estimation, assume for simplicity that $\pi_2(r)$ has nonzero slope at $r = r^-$ and $r = r^+$.[37] Let

$$\hat{\pi}_1(r) = (r - c_0)\big(1 - \widehat{F}_U^+(r)\big),$$
$$\hat{\pi}_2(r) = (r - c_0)\big(1 - \widehat{F}_U^-(r)\big),$$
$$\hat{\pi}_1^* = \sup_r \hat{\pi}_1(r),$$
$$\hat{r}_1^* = \arg\sup_r \hat{\pi}_1(r)$$

and define the correspondence $\pi_2^c(\cdot)$ by

$$\big\{\pi \in \pi_2^c(r)\big\} \quad \Longleftrightarrow \quad \Big\{\lim_{r' \downarrow r} \hat{\pi}_2(r') \leqslant \pi \leqslant \lim_{r' \uparrow r} \hat{\pi}_2(r')\Big\}.$$

[37] Haile and Tamer (2003) provide estimators that do not require this assumption.

This defines a smooth sample analog of $\pi_2(\cdot)$ that can be used to define consistent estimators of r^- and r^+:

$$\hat{r}^- = \sup\{r < \hat{r}_1^*\colon\ \pi = \hat{\pi}_1^* \text{ for some } \pi \in \pi_2^c(r)\},$$
$$\hat{r}^+ = \inf\{r > \hat{r}_1^*\colon\ \pi = \hat{\pi}_1^* \text{ for some } \pi \in \pi_2^c(r)\}.$$

4.3.3. Asymmetric and affiliated private values

In principle, the applicability of Assumptions 4.1 and 4.2 is not limited to environments with symmetric independent private values. Haile and Tamer (2001) have explored extensions to models of asymmetric and/or affiliated private values. While it is encouraging that any restrictions at all on the joint distribution $F_{\mathbf{U}}(\cdot)$ can be obtained without the assumption of independence that was required for identification in the button auction model, in practice the bounds one can obtain without the independence assumption are likely to be quite wide. Intuitively, when one observes only bounds on realizations of random variables, it is difficult to learn much about their correlation structure. Of course, without knowledge of the correlation structure, a number of important positive and normative questions cannot be answered.[38] Thus, while the bounds approach provides a way of addressing concerns about the appropriateness of the standard button auction model, it may provide little help in environments in which the button auction model itself is unidentified.

5. Specification testing

Identification of the models discussed above relies on behavioral assumptions and on assumptions about the underlying demand and information structure. Obviously, then, the choice of model is important. In some environments there are overidentifying restrictions that can be used to test some assumptions while maintaining others. Several testing approaches have been described in the literature to date, although so far there has been little attention to development of formal statistical tests.

[38] Optimal auction design with correlated valuations is much more complex than in the IPV case, requiring precise information about the underlying correlation structure [cf., Crémer and McLean (1988) and McAfee and Reny (1992)]. Quint (2004) has shown that even the simpler question of the optimal reserve price cannot be addressed with a bounds approach in such an environment. In particular, for any reserve price $r \geqslant u_0$ and any distribution of bids, there exists an underlying joint distribution of valuations consistent with these bids and Assumptions 1 and 2 such that r is the optimal reserve price. Hence, no restriction on the optimal reserve price can be obtained from nondegenerate bounds on the joint distribution of valuations.

5.1. Theoretical restrictions in first-price auction models

We first consider restrictions imposed by equilibrium bidding in first-price auctions.[39] Recall that a model is testable if there exists some joint distribution of observables that cannot be rationalized by the model. It is then natural to ask what set of distributions *can* be rationalized. Here we provide two results for the affiliated private values (APV) framework.[40] The first gives necessary conditions for a distribution of bids to be rationalized by equilibrium behavior, while the second gives necessary and sufficient conditions in two special cases: symmetric affiliated private values, or independent private values (IPV).

THEOREM 5.1. *Consider the APV first-price auction with fixed* $\mathcal{N}$. *Necessary conditions for* $G_{\mathbf{B}}(\cdot;\mathcal{N})$ *to be rationalized by equilibrium bidding are*

(a) $(B_1,\ldots,B_n)$ *are affiliated*;
(b) *for each* i, $\xi_i(\cdot,\mathcal{N})$ *is continuous and strictly increasing on* $\operatorname{supp}[B_i]$;
(c) $\operatorname{supp}[\xi_1(B_1,\mathcal{N})]\times\cdots\times\operatorname{supp}[\xi_n(B_n,\mathcal{N})]$ *is a convex, compact set, and on this set the joint distribution* $G_{\mathbf{B}}(\xi_1^{-1}(u_1,\mathcal{N}),\ldots,\xi_n^{-1}(u_n,\mathcal{N}))$ *is absolutely continuous* (*with respect to the Lebesgue measure*) *as a function of* $\mathbf{u}$, *with a strictly positive density*;
(d) $\underline{b}_i=\underline{b}$ *for all* i, *and* $\xi_i(\underline{b},\mathcal{N})=\underline{b}$ *for all* $i\in\mathcal{N}$;
(e) $\xi_i(\bar{b}_i,\mathcal{N})=\xi_j(\bar{b}_j,\mathcal{N})$ *for all* i, $j\in\mathcal{N}$; *and*
(f) *for each* i, $\operatorname{supp}[B_i]\subseteq\operatorname{supp}[\max_{j\in\mathcal{N}\setminus i}B_j]$, *and* $\operatorname{supp}[\max_{j\in\mathcal{N}\setminus i}B_j]$ *is convex.*

PROOF. Given strictly increasing bidding strategies and affiliated private values with an atomless type distribution, affiliation of bids and strict monotonicity of $\xi_i(\cdot,\mathcal{N})$ on $\operatorname{supp}[B_i]$ follow directly. Continuity of $\xi_i(\cdot,\mathcal{N})$ follows from strict monotonicity of the bidding strategies together with the assumption that $\operatorname{supp}F_{U_i}(\cdot)$ is convex. Since Assumption 2.1 requires that $F_{\mathbf{U}}(\cdot)$ have a strictly positive joint density on a compact convex set, the relationship between $F_{\mathbf{U}}(\cdot)$ and $G_{\mathbf{B}}(\cdot)$ given by (3.2) implies that (c) must hold. The assumption that $\operatorname{supp}F_{U_i}(\cdot)$ does not vary with i, together with the equilibrium conditions $\max(r,\underline{u})=\underline{b}_i$ and $\beta_i(\max(r,\underline{u}))=\max(r,\underline{u})$ for each i, imply (d) and (e). The necessity of $\operatorname{supp}[B_i]\subseteq\operatorname{supp}[\max_{j\in\mathcal{N}\setminus i}B_j]$ follows from (d) and the fact that when bidding against opponents who use strictly increasing strategies it is never optimal for bidder i to bid more than the minimum necessary to win with a particular probability. The same logic implies that $\operatorname{supp}[\max_{j\in\mathcal{N}\setminus i}B_j]$ is convex, since

[39] Recall that we maintain Assumption 2.1, restricting the primitives of the model, and that we focus on equilibria in strictly increasing strategies. Theorem 2.1 guarantees that such an equilibrium exists for the APV model. If bidders are symmetric, Theorem 2.1 implies that there is a unique equilibrium in the class of equilibria in nondecreasing strategies. When bidders are asymmetric, we do not have a uniqueness result.

[40] Guerre, Perrigne and Vuong (2000) gave a similar result for the symmetric independent private values model. See also Li, Perrigne and Vuong (2002).

no bidder $j \in \mathcal{N}\backslash i$ would find it optimal to place a bid at the upper boundary of a gap in this support. □

Note that we do not provide sufficient conditions for $G_{\mathbf{B}}(\cdot)$ to be rationalized in the APV model with asymmetric bidders, since a full equilibrium characterization is not available for that case. However, in the special cases of IPV or symmetric bidders when valuations have a continuously differentiable density, Theorem 2.1 implies that there is a unique equilibrium, which has strictly increasing and differentiable strategies and the same support for the equilibrium bids of all bidders. When bidders are symmetric, the unique equilibrium is symmetric. In these settings we have necessary and sufficient conditions for a bidding distribution to be rationalized. The statement of the conditions of Theorem 5.1 can then also be simplified somewhat, exploiting differentiability of strategies.

THEOREM 5.2. *Consider the APV first-price auction with fixed* $\mathcal{N}$. *Assume that* $f_{\mathbf{U}}(\cdot)$ *is continuously differentiable and suppose, further, that either*

(i) $(U_1, \ldots, U_n)$ *are mutually independent or*
(ii) *bidders are symmetric.*

Necessary and sufficient conditions for $G_{\mathbf{B}}(\cdot; \mathcal{N})$ *to be rationalized by equilibrium bidding are*:

(a) $(B_1, \ldots, B_n)$ *are affiliated, and in case* (i) *they are independent, while in case* (ii) *they are exchangeable*;
(b) *for each* i, $\xi_i(\cdot, \mathcal{N})$ *is differentiable and strictly increasing on* $\text{supp}[B_i]$;
(c) $G_{\mathbf{B}}(\xi_1^{-1}(u_1, \mathcal{N}), \ldots, \xi_n^{-1}(u_n, \mathcal{N}))$ *is absolutely continuous* (*with respect to the Lebesgue measure*) *as a function of* $\mathbf{u}$, *with a positive continuously differentiable density on* $\text{supp}[\xi_1(B_1, \mathcal{N})] \times \cdots \times \text{supp}[\xi_n(B_n, \mathcal{N})]$ *and zero density elsewhere*;
(d) $\xi_i(\underline{b}, \mathcal{N}) = \underline{b}$ *for all* $i \in \mathcal{N}$;
(e) $\xi_i(\overline{\overline{b}}_i, \mathcal{N}) = \xi_j(\overline{\overline{b}}_j, \mathcal{N})$ *for all* $i, j \in \mathcal{N}$; *and*
(f) $\text{supp}[B_i]$ *is convex, compact, and the same for all* i.

PROOF. Given strictly increasing, differentiable bidding strategies and the conditions on the $F_{\mathbf{U}}(\cdot)$, affiliation of bids and the relevant independence and symmetry conditions follow directly. For condition (f), equal supports is necessary by Theorem 2.2 in case (i) and by symmetry in case (ii). Convex, compact support follows because bidding strategies are strictly increasing, continuous functions of random variables with convex, compact support. Condition (b) is necessary because differentiability of $\xi_i(\cdot, \mathcal{N})$ is equivalent to differentiability of $\xi_i^{-1}(\cdot, \mathcal{N})$ (since $\xi_i(\cdot, \mathcal{N})$ is strictly increasing), with the latter equal to the equilibrium bidding strategy under the assumptions of the model. Conditions (d) and (e) are necessary following the arguments in Theorem 5.1. Recall that Assumption 2.1 requires that $F_{\mathbf{U}}(\cdot)$ have a strictly positive joint density on a compact, convex set, and that we have assumed that it has a differentiable density. The set $\text{supp}[\xi_1(B_1, \mathcal{N})] \times \cdots \times \text{supp}[\xi_n(B_n, \mathcal{N})]$ is the support of valuations implied by the

model, and it is convex and compact by (f) and differentiability of $\xi_i(\cdot)$. The relationship between $F_{\mathbf{U}}(\cdot)$ and $G_{\mathbf{B}}(\cdot)$ specified by (3.2) then implies (c).

To see that the stated conditions are sufficient for $G_{\mathbf{B}}(\cdot;\mathcal{N})$ to be rationalized, observe that they ensure that $\xi_i^{-1}(\cdot)$ is well defined, differentiable, and strictly increasing on $\text{supp}[\xi_i(B_i)]$ for each i, and that in symmetric models it is the same for each i, so that the expression for $F_{\mathbf{U}}(\cdot)$ in (3.2) is well defined and satisfies the relevant affiliation, independence, symmetry, and differentiability conditions. The conditions guarantee that the implied $F_{\mathbf{U}}(\cdot)$ has a support that is convex and compact; that it has a strictly positive, continuously differentiable density on this support; and that the support is the same for all bidders. The definition of $\xi_i(\cdot)$ implies that if each bidder i uses the bidding strategy $\xi_i^{-1}(\cdot)$, his first-order condition for optimality is satisfied. Under independence or symmetry, bidder payoffs satisfy a single crossing property: for any fixed monotone strategies by opposing bidders, a higher realized valuation u_i leads to a higher marginal return to increasing one's bid. Standard results from the literature on auctions and mechanism design [see, e.g., Fudenberg and Tirole (1991)] imply when the single crossing property holds, local optimality of a bid (i.e., first-order conditions hold) together with monotonicity of the bidding strategy are necessary and sufficient for global optimality of the strategy. Thus the strategies $\{\xi_i^{-1}(\cdot)\}_{i\in\mathcal{N}}$ form an equilibrium. □

The importance of a result providing sufficient conditions for a bid distribution to be rationalized by equilibrium behavior should not be underappreciated. Without such a result, one would have no way of ensuring that the interpretation of bids based on the first-order conditions is valid. In particular, for an observed bid b_i and an implied valuation $u_i = \xi_i(b_i)$, there would be no guarantee that b_i was actually an equilibrium bid for a bidder with valuation u_i. Verifying that the observed bids actually can be rationalized by equilibrium behavior is analogous to verifying second-order conditions for optimality: only when such sufficient conditions are verified can we be sure that the mappings (forward or inverse) provided by the first-order conditions relate valuations to optimal (best-response) bids. Theorem 5.2 can then be used in two ways. First, in an application one can attempt to verify that sufficient conditions for bid data to be consistent with the assumptions of the model are satisfied. Second, the necessary conditions suggest specification tests, which we discuss further in the following section.

We note that it is possible to generalize the overall empirical approach to the case where condition (e) above fails by relaxing the assumption that $\text{supp}[U_i]$ is the same for all bidders i. The latter assumption is typically maintained in the literature [see, e.g. Campo, Perrigne and Vuong (2003)]. In independent private values models, it ensures that $\text{supp}[B_i]$ is the same for all bidders i [Lebrun (1999)] and that the equilibrium is unique [Lebrun (1999), Bajari (2001)]. However, plausible specifications of primitives would lead to distributions of bids that violate condition (e), so it may be useful to relax that assumption in practice.

Let us briefly consider some examples. Consider maintaining the assumption that $\inf[\text{supp}[U_i]]$ is the same for all i, but allow $\overline{u}_i = \sup[\text{supp}[U_i]]$ to vary with i. For affili-

ated private values models, there exists an equilibrium in nondecreasing strategies where $\text{supp}[B_i] \subseteq \text{supp}[\max_{j \neq i} B_j]$, and these supports are convex and compact; in addition, in any equilibrium (mixed or pure) strategies must be strictly increasing (i.e., separating) [Maskin and Riley (2003), McAdams (2007)]. Thus, the distributions of valuations can be identified using (3.1). For example, suppose $n = 2$, U_1 is uniformly distributed on $[0, 3/2]$, and U_2 is independent of U_1, with distribution $F_{U_2}(u_2) = (1/4)u_2^2$ on the support $[0, 2]$. Then, it can be shown that $\text{supp}[B_i] = [0, 1]$ for $i \in \{1, 2\}$, and that for $b \in [0, 1]$, $G_{B_1}(b) = b$ and $G_{B_2}(b) = b^2$. With these bid distributions, $g_{M_1}(1; \mathcal{N}) = 1$ while $g_{M_2}(1; \mathcal{N}) = 2$, violating the boundary condition (e). However, in this example, the $F_{U_i}(\cdot)$ each can be identified if we expand the set of permissible distributions of valuations to allow supports that vary across bidders.

If both $\inf[\text{supp}[U_i]]$ and $\sup[\text{supp}[U_i]]$ vary with i, then it is possible that some bidders never win in equilibrium. For example, if there are two bidders in an IPV auction, and $\text{supp}[U_1] = [0, 1]$ while $\text{supp}[U_2] = [100, 101]$, in equilibrium $B_2 = 1$ with probability 1, while $B_1 \leqslant 1$ Maskin and Riley (2000a). Clearly, very little can be said about the distribution of U_2 in this case. Despite the possibility of degenerate equilibria like this, Maskin and Riley (2000b) show that the distribution of winning bids, $G_B^{(n:n)}(\cdot)$, is continuous on its support. This implies that a mass point in $G_B^{(n:n)}(\cdot)$ such as the one in the latter example can only occur either (i) at the bottom of the support if the support of winning bids is nondegenerate, or (ii) if the support of winning bids is degenerate. Thus, outside of cases (i) and (ii), the equilibrium must be in strictly increasing strategies on $\text{supp}[B^{(n:n)}]$, so that it will be possible to recover the distribution of bidders' valuations on the pre-image of the interior of $\text{supp}[B^{(n:n)}]$ using (3.1). This would lead to a partial identification result.

For the remainder of the chapter, we follow the existing literature and maintain the assumption that valuation distributions have the same support, while noting that many of the results generalize.

5.2. *Testing monotonicity of bid functions and boundary conditions*

Here we consider two possible types of tests based on Theorem 5.1. Guerre, Perrigne and Vuong (2000) have suggested a specification test based on the observation that the right-hand side of bidder i's first-order condition (2.4), i.e.,

$$\xi_i(b_i, \mathcal{N}) \equiv b_i + \frac{G_{M_i|B_i}(b_i|b_i; \mathcal{N})}{g_{M_i|B_i}(b_i|b_i; \mathcal{N})}$$

is the inverse of his equilibrium bidding strategy. This is true for private and common value auctions.[41] Since bidding strategies must be strictly increasing, so must $\xi_i(\cdot, \mathcal{N})$.

[41] One way to see this in the common values case is to note that one possible normalization of signals sets $X_i = v_i(X_i, X_i; \mathcal{N})$, so that the first order condition may still be directly interpreted as giving bidder i's inverse bidding strategy.

Although testing monotonicity of $\xi_i(\cdot, \mathcal{N})$ is conceptually straightforward, no formal statistical test has been developed for this problem. Existing tests of monotonicity in the statistics literature[42] are not directly applicable due to the fact that realizations of the random variables $\xi_i(B_i, \mathcal{N})$ are estimated rather than observed directly. In applications, researchers often find few (if any) violations of strict monotonicity [e.g., Hendricks, Pinkse and Porter (2003), Haile, Hong and Shum (2003)], in which case no formal test would reject. Nonetheless, formal tests could be valuable.

Two things about such a test should be noted, however. First, the alternative hypothesis is simply that some component of the specification is incorrect. A failure of monotonicity may indicate the presence of unobserved heterogeneity, risk aversion, nonequilibrium bidding behavior, or violation of some other maintained assumption. In general, testing one assumption will require maintaining others, so many of the other specification tests discussed below will share this limitation. Second, no test of this hypothesis will be consistent against all violations of the maintained assumptions. In particular, one can easily construct examples in which one or more maintained assumptions are violated, but monotonicity of $\xi_i(\cdot, \mathcal{N})$ still holds.

Another potential specification test is based on the boundary condition (e) from Theorem 5.1. This restriction can be simplified in the case of the IPV model. Let the common support of the bid distribution be denoted $\text{supp}[B_i] = [\underline{b}, \overline{b}]$. Then, the boundary condition requires

$$g_{M_i}(\overline{b}; \mathcal{N}) = g_{M_j}(\overline{b}; \mathcal{N}) \quad [\forall i, j \in \mathcal{N}] \tag{5.1}$$

which is a testable restriction.

5.3. Multi-sample tests with exogenous variation in participation

Athey and Haile (2002) discuss a different principle for specification testing that can be used in both first-price and ascending auctions whenever (a) there is exogenous variation in the number of bidders, and (b) the underlying model is identified with a fixed number of bidders. For simplicity, consider the case of symmetric bidders, although the same principle applies to asymmetric settings. Let $\widehat{F}_{\mathbf{U}}(\mathbf{u}; n)$ denote a consistent estimator of $F_{\mathbf{U}}(\mathbf{u})$ obtained using data from n-bidder auctions. With exogenous variation in the number of bidders, for $n' \neq n$, $\widehat{F}_{\mathbf{U}}(\mathbf{u}; n)$ should equal $\widehat{F}_{\mathbf{U}}(\mathbf{u}; n')$ up to sampling error. Hence a test of the null hypothesis of equal distributions provides a specification test.

While testing equality of distributions is a standard problem [e.g., McFadden (1989)], complications arise both in ascending and first-price auctions. In an ascending auction, the complication is the fact that identification relies on mappings between distributions of order statistics and the underlying marginal distributions (Theorem 4.1). Hence, asymptotic distributions of test statistics must account for this transformation of the data.

[42] See, e.g., Bowman, Jones and Gijbels (1998), Gijbels et al. (2000), or Hall and Heckman (2000).

In a first-price auction, the complications are more challenging, arising from the fact that valuations are estimated rather than observed directly. This first-stage nonparametric estimation of $F_{\mathbf{U}}(\cdot)$ introduces nontrivial complications to the asymptotic theory needed for inference. Haile, Hong and Shum (2003) develop several formal tests applicable when all bids are observed, based on comparisons of different estimates of the marginal distribution $F_U(\cdot)$ obtained from auctions with different number of bidders.[43] In models identified with partially observed bids, similar tests may be applicable, although this has not yet been explored.

5.4. Multi-sample tests with multiple order statistics

In IPV settings, a variation on the type of testing approach above may be available without exogenous variation in participation. In an IPV auction each marginal distribution $F_{U_i}(\cdot)$ is identified from observation of the transaction price (and bidder identities if the environment is asymmetric) in both ascending and first-price auctions. Athey and Haile (2002) have shown that observation of any other order statistic $B^{(j:n)}$ can be substituted for observability of the transaction price – in a symmetric environment, for example, this follows from (4.1). When two or more order statistics (e.g., the top two bids) are observed, the estimates of $F_{U_i}(\cdot)$ implied by each of these should be identical up to sampling error.

5.5. Direct tests of exchangeability or independence

There are other potential approaches to specification testing when bidders are assumed to be symmetric or types are assumed independent. With symmetric bidders, the joint distribution of bidder valuations is exchangeable and each bid $B_i = \beta(U_i; n)$. Hence, the joint distribution of bids must also be exchangeable. When bidder identities are observed, there are several ways to approach testing such a hypothesis. One is to test exchangeability of the bids (or subsets of bids) directly. Nonparametric tests from the statistics literature may be directly applicable. For example, Romano (1988, 1989) suggests tests based on the supremum distance between the values of a multivariate CDF evaluated at permutations of its arguments.

One implication of exchangeability is equality of marginal distributions. For example, in a symmetric model, any subset of bidders should have bids governed by the same marginal distribution as those of another subset of bidders. A standard Kolmogorov–Smirnov test of equal distributions could then be applied.

Alternative tests may be useful when covariates are available and additional structure is assumed. Suppose, for example, that valuations are assumed to have the structure

$$U_{it} = h(Z_{1t}, Z_{2i}, \mathbf{Z}_{2(-i)}, A_{it})$$

[43] They focus on tests of the private values hypothesis. However, their tests, which are based on comparisons of the empirical distributions of pseudo-values for auctions with different numbers of bidders, could be directly applied.

where Z_{1t} is an auction-specific covariate, Z_{2i} is a bidder-specific covariate, $\mathbf{Z}_{2(-i)}$ denotes the bidder-specific covariates of i's opponents, and A_{it} is a private idiosyncratic factor. The restriction to scalar covariates is only for expositional simplicity. Assume further that the conditional distribution function $F_{\mathbf{A}}(A_1, \ldots, A_n | Z_1, Z_{21}, \ldots, Z_{2n})$ is exchangeable in the indices $(1, \ldots, n)$. Loosely speaking, with this structure, all bidder valuations are affected in the same way by covariates. In particular, the distribution of bidder i's valuation conditional on $(Z_1, Z_{2i}, \mathbf{Z}_{2(-i)})$ is the same for all i. Since bids are equal to valuations in an ascending auction, this can be tested, for example, by examining coefficient estimates in a regression of bids on covariates (auction- and bidder-specific) interacted with bidder dummies (or indicators for different "classes" of bidders).

In a first-price auction, the structure above implies that the distribution of $\max_{j \neq i} B_j$ is the same for all i conditional on $(Z_1, Z_{2i}, \mathbf{Z}_{2(-i)})$. Hence, the distribution of i's bids should depend only on $(Z_1, Z_{2i}, \mathbf{Z}_{2(-i)})$, not on the index i itself. This may again be evaluated in a regression. Bajari and Ye (2003) apply these regression-based approaches in their analysis of highway construction contracts [see also Porter and Zona (1993, 1999)].

Note that similar restrictions will hold in a common values model, where it is the joint distribution of the random variables

$$v_i(X_i, X_i, \mathcal{N})$$

that must be exchangeable. As we will see below, this distribution will often be identified in a common values model, even though $F_{\mathbf{U},\mathbf{X}}(\cdot)$ is not identified. Hence, specification testing may be possible even for under-identified models.

Another direct approach to specification testing is applicable in first-price auctions in the widely used independent private values model (symmetric or asymmetric). Since each B_i is a measurable function of U_i, bids must also be independent. In a first-price sealed-bid auction in which all bids are observed, one can directly test this restriction using standard nonparametric tests [Guerre, Perrigne and Vuong (2000)]. For example, Romano (1988, 1989) suggests tests based on the supremum distance between an estimated joint distribution and the joint distribution obtained as the product of the estimated underlying marginal distributions.[44] In practice, it is typical to assume that valuations, and thus bids, are independent conditional on a set of auction-specific and perhaps bidder-specific covariates. Su and White (2003) propose a testing approach that may then be applicable. An alternative is to test for correlation of residuals from a regression of bids on bidder-specific or auction-specific covariates. Bajari and Ye (2003) do this in their analysis of highway construction procurement auctions. In an ascending auction, the problem of partially observed bids appears to make direct testing impossible (recall, however, the indirect tests discussed in Section 5.4).

[44] Other tests of the hypothesis that bids are uncorrelated (an implication of independence) could also be applied. See, e.g., Chapter 8 of Hollander and Wolfe (1999).

6. Extensions of the basic results

6.1. Auction heterogeneity

6.1.1. Observed auction heterogeneity

In practice, one rarely has access to data from auctions of identical objects. For example, the goods for sale at each auction often differ in observable characteristics, and we may expect distributions of valuations to shift with these observables. All of the identification results above hold in the presence of auction-specific covariates. In particular, the previous discussion can be reinterpreted as being conditioned on a given realization of the covariate values. To make this concrete, let $\mathbf{Z}$ be a vector of auction covariates. We extend the notation defined above to condition on $\mathbf{Z}$ by defining $\beta_i(\cdot;\mathcal{N},\mathbf{Z})$, $F_{\mathbf{U}}(\cdot|\mathbf{Z})$, $G_{M_i|B_i}(b|b;\mathcal{N},\mathbf{Z})$ and $g_{M_i|B_i}(b|b;\mathcal{N},\mathbf{Z})$, etc. Assuming all auction-specific heterogeneity is captured by $\mathbf{Z}$, in a first-price auction the first-order condition for bidder i at auction t becomes

$$u_{it} = b_{it} + \frac{G_{M_i|B_i}(b_{it}|b_{it};\mathcal{N},\mathbf{z}_t)}{g_{M_i|B_i}(b_{it}|b_{it};\mathcal{N},\mathbf{z}_t)} \tag{6.1}$$

which uniquely determines $F_{\mathbf{U}}(\cdot|\mathbf{z}_t)$ in the affiliated private values model when all bids and bidder identities are observable. In an ascending auction with private values that are independent conditional on $\mathbf{Z}_t$, one can use the conditional distribution of transaction prices $F_U^{(n-1:n)}(\cdot|\mathbf{z}_t)$ for any given value of $\mathbf{z}_t$ to uniquely determine $F_{\mathbf{U}}(\cdot|\mathbf{z}_t)$ through Equation (4.4).

The nonparametric estimation methods discussed above can also be extended, for example by using standard kernel smoothing over covariates. Guerre, Perrigne and Vuong (2000) discuss details of such an approach for the case of a first-price auction with symmetric independent private values, and this approach is easily extended to the other models. This type of approach has been applied to ascending auctions by Haile and Tamer (2003).

Unless the dimensionality of the covariates is fairly small relative to the sample size, however, a fully general nonparametric estimation approach may not be practical. One alternative suggested by Haile, Hong and Shum (2003) exploits the observation that additive (or multiplicative) separability is preserved by equilibrium bidding.[45] In particular, suppose that in an auction with characteristics $\mathbf{z}_t$ valuations are given by

$$u_{it} = \Gamma(\mathbf{z}_t) + a_{it} \tag{6.2}$$

for some (possibly unknown) function $\Gamma(\cdot)$, with the bidder-specific private information A_{it} independent of $\mathbf{Z}_t$. Then, if we let $\mathbf{z}_0$ be such that[46]

[45] This approach has been applied by Krasnokutskaya (2004), Bajari, Houghton and Tadelis (2004), and Shneyerov (2005).

[46] We assume for simplicity that such a $\mathbf{z}_0$ exists. If it does not, the argument extends but with more cumbersome notation.

$$\Gamma(\mathbf{z}_0) = 0 \tag{6.3}$$

equilibrium bidding also follows the additively separable structure (an analogous result applies in the case of multiplicative separability)

$$\beta_i(u_i; \mathcal{N}, \mathbf{z}) = \Gamma(\mathbf{z}) + \beta_i(u_i; \mathcal{N}, \mathbf{z}_0). \tag{6.4}$$

Proving this is trivial in an ascending auction, where the bid function is the identity function. For a first-price sealed-bid auction, let

$$\breve{\beta}_i(a_i, \mathbf{z}; \mathcal{N}) \equiv \beta_i\big(a_i + \Gamma(\mathbf{z}); \mathcal{N}, \mathbf{z}\big)$$

so that under (6.2) a bidder's first-order condition can be written

$$a_{it} + \Gamma(\mathbf{z}_t) = \breve{\beta}_i(a_{it}, \mathbf{z}_t; \mathcal{N}) + \frac{G_{M_i|B_i}(\breve{\beta}_i(a_{it}, \mathbf{z}_t; \mathcal{N})|\breve{\beta}_i(a_{it}, \mathbf{z}_t; \mathcal{N}); \mathcal{N}, \mathbf{z}_t)}{g_{M_i|B_i}(\breve{\beta}_i(a_{it}, \mathbf{z}_t; \mathcal{N})|\breve{\beta}_i(a_{it}, \mathbf{z}_t; \mathcal{N}); \mathcal{N}, \mathbf{z}_t)}. \tag{6.5}$$

Note that the events $\{\breve{\beta}_i(A_i, \mathbf{z}; \mathcal{N}) = \breve{\beta}_i(a_i, \mathbf{z}; \mathcal{N})\}$ and $\{\breve{\beta}_i(A_i, \mathbf{z}_0; \mathcal{N}) = \breve{\beta}_i(a_i, \mathbf{z}_0; \mathcal{N})\}$ are equivalent for any $\mathbf{z}$. Under (6.4), the events $\{\breve{\beta}_j(A_j, \mathbf{z}; \mathcal{N}) = \breve{\beta}_i(a_i, \mathbf{z}; \mathcal{N})\}$ and $\{\breve{\beta}_j(A_j, \mathbf{z}_0; \mathcal{N}) = \breve{\beta}_i(a_i, \mathbf{z}_0; \mathcal{N})\}$ are equivalent for $j \neq i$, so the expression

$$\frac{G_{M_i|B_i}(\breve{\beta}_i(a_{it}, \mathbf{z}_t; \mathcal{N})|\breve{\beta}_i(a_{it}, \mathbf{z}_t; \mathcal{N}); \mathcal{N}, \mathbf{z}_t)}{g_{M_i|B_i}(\breve{\beta}_i(a_{it}, \mathbf{z}_t; \mathcal{N})|\breve{\beta}_i(a_{it}, \mathbf{z}_t; \mathcal{N}); \mathcal{N}, \mathbf{z}_t)}$$

on the right-hand side of (6.5) is invariant to $\mathbf{z}_t$. Hence, (6.4) guarantees that (6.5) is satisfied for all $\mathbf{z}_t$ whenever it is for $\mathbf{z}_t = \mathbf{z}_0$.

This preservation of additive separability is useful because it implies that the effects of covariates on valuations can be controlled for using a regression of bids on covariates. In particular, we can write

$$b_{it} = \alpha(\mathcal{N}_t) + \Gamma(\mathbf{z}_t) + \epsilon_{it}, \tag{6.6}$$

where $\alpha(\mathcal{N}_t)$ is an intercept specific to auctions in which the set of bidders is $\mathcal{N}_t$ (in a symmetric environment, this can be $\alpha(n_t)$) and $\epsilon_{it} \equiv \beta_i(u_{it}; \mathcal{N}_t, \mathbf{z}_0) - \alpha(\mathcal{N}_t)$ has mean zero conditional on $\mathbf{z}_t$. Both $\alpha(\mathcal{N}_t)$ and $\Gamma(\mathbf{z}_t)$ are then identified from observation of bids, $\mathcal{N}_t$, and $\mathbf{z}_t$; indeed, they can be estimated consistently using standard regression techniques.

Let $\widehat{\Gamma}(\mathbf{z}_t)$ denote a consistent estimate of $\Gamma(\mathbf{z}_t)$. Then $b_{it} - \widehat{\Gamma}(\mathbf{z}_t)$ provides a consistent estimate of $\beta_i(u_{it}; \mathcal{N}_t, \mathbf{z}_0)$, i.e., the bid i *would have submitted* in auction t if $\mathbf{Z}_t$ were equal to $\mathbf{z}_0$. Of course, a sample of bids from auctions with the same value of $\mathbf{Z}$ of is exactly what we would like to have. Estimation of (6.6) provides an approach for "homogenization" of the bid data by replacing each b_{it} with

$$b_{it}^h = b_{it} - \widehat{\Gamma}(\mathbf{z}_t).$$

These homogenized bids can then be used to consistently estimate the underlying distribution of valuations $F_{\mathbf{U}}(\cdot; \mathbf{z}_0)$ using the methods described in the previous sections;

i.e., with $\Gamma(\cdot)$ known, $F_{\mathbf{U}}(\cdot;\mathbf{z}_0)$ is identified through (6.1). Finally, since (6.3) and (6.2) imply

$$\Pr(U_{1t}\leqslant u_1,\ldots,U_{nt}\leqslant u_n)=F_{\mathbf{U}}\big(u_1-\Gamma(\mathbf{z}_{it}),\ldots,u_n-\Gamma(\mathbf{z}_{it});\mathbf{z}_0\big),$$

$F_{\mathbf{U}}(\cdot;\mathbf{z})$ is then identified for all $\mathbf{z}$ in the support of the auction covariates. As usual, in a first-price auction, equilibrium bidding implies that the distribution of the mean-zero ϵ_{it} will vary with $\mathcal{N}_t$. So the "second stage" of estimating the joint distribution $F_{\mathbf{U}}(\cdot;\mathbf{z}_0)$ must be done separately for each $\mathcal{N}_t$.

The number of observations available for the first-stage regression of bids on covariates is $\sum_n nT_n$, which is often quite large. Hence, a nonparametric or flexible parametric specification of $\Gamma(\cdot)$ will be feasible in data sets of reasonable size. Assuming that $\Gamma(\cdot)$ is known up to a finite parameter vector has an advantage for some purposes in that estimates from the first stage will converge at the parametric rate, leaving the asymptotic distribution of nonparametric estimators applied to the homogenized sample unaffected. Note that the function $\Gamma(\cdot)$, which characterizes the effects of covariates on valuations, is sometimes of direct interest itself. Equation (6.4) implies that one can estimate this primitive directly with a regression of bids on covariates. Bajari, Houghton and Tadelis (2004), for example, have recently exploited this observation to investigate the importance of renegotiation costs in procurement auctions.

This approach preserves the fully nonparametric specification of the idiosyncratic component of bidders' private values and allows direct inference (through the first-stage estimates) on the way observables affect valuations. However, it places a strong restriction on the way observables enter. An alternative nonparametric approach is to use series or sieves [e.g., Chen (2007)], approximating the bid distribution with a sequence of parametric models. In a given data set this will amount to assuming a flexible parametric model, and one might also take such an approach directly. For example, in an ascending auction with symmetric independent private values, one might specify the conditional distribution $F_U^{(n-1:n)}(u|\mathbf{z})$ as a finite mixture of parametric distributions. Letting $H(\cdot;\boldsymbol{\gamma})$ be a parameterized distribution function, the distribution of the transaction price could be specified as

$$F_U^{(n-1:n)}(u|\mathbf{z},\boldsymbol{\theta},J)=\frac{1}{\sum_{j=1}^{J}\omega(\mathbf{z};\boldsymbol{\theta}_j)}\sum_{j=1}^{J}\omega(\mathbf{z};\boldsymbol{\theta}_j)H\big(u;\boldsymbol{\gamma}(\mathbf{z};\boldsymbol{\theta}_j)\big)\tag{6.7}$$

given parametric specifications of the functions $\boldsymbol{\gamma}(\cdot)$ and $\omega(\cdot)$. Given an estimate $\hat{\boldsymbol{\theta}}$ of the parameter vector $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_J)$, Equation (4.9) implies that

$$\widehat{F}_U(u|\mathbf{z})\equiv\phi\big(F_U^{(n-1:n)}(u|\mathbf{z};\hat{\boldsymbol{\theta}},J);n-1,n\big)\tag{6.8}$$

would provide a consistent estimator of $F_U(u|\mathbf{z})$ under (6.7).

In a first-price sealed-bid auction, a similar approach might be applied. For a given set of bidders $\mathcal{N}$, the conditional distribution $G_{M_i|B_i}(m_i|b_i;\mathcal{N},\mathbf{z})$ could be assumed to

have the parametric form

$$G_{M_i|B_i}(m_i|b_i; \mathcal{N}, \mathbf{z}, \boldsymbol{\theta}, J) = \sum_{j=1}^{J} \frac{\omega(b_i, \mathbf{z}; \boldsymbol{\theta}_j)}{\sum_{j'=1}^{J} \omega(b_i, \mathbf{z}; \boldsymbol{\theta}_{j'})} H\big(m; \boldsymbol{\gamma}(b_i, \mathbf{z}; \boldsymbol{\theta}_j)\big) \tag{6.9}$$

providing a flexible parametric first step of the two-step estimation procedure discussed in Section 3.2. This kind of specification allows the distribution of bids to vary with auction covariates in richer ways than is allowed by the "homogenization" approach described above. This flexibility comes at the price of placing *a priori* structure on the distributions of bids and valuations. Of course, some approximation must always be used in a finite sample, and a finite mixture may perform well in practice. Note that here the effects of covariates on valuations, which are often of primary interest, would be obtained indirectly, through (6.8) or through (6.1) and (6.9).

6.1.2. Unobserved auction heterogeneity

In many applications one may suspect that there are factors affecting bidder valuations that are common knowledge among bidders but unobserved by the econometrician. For example, suppose valuations are given by the equation

$$U_i = V_0 + A_i. \tag{6.10}$$

Even if the idiosyncratic components $A_1, \ldots, A_n$ are i.i.d., the valuations $U_1, \ldots, U_n$ will be correlated unconditional on V_0 (they will be affiliated if the densities $f_{A_i}(\cdot)$ are log-concave). When bidders' information consists only of their valuations u_i, not the individual components v_0 and a_i, this merely provides one motivation for an affiliated private values model. When bidders observe both v_0 and a_i, however, the situation can be more complicated.

As noted above, information regarding an auction that is common knowledge among the bidders creates no problem for the characterization of equilibrium bidding strategies – the theory can be thought of as holding for each value of the public information. However, for empirical work, difficulties can arise when the econometrician is unable to condition on all the information that is public to bidders.

There are at least three issues that arise in the presence of unobserved heterogeneity. The first is whether unobserved heterogeneity is empirically distinguishable from other structures that introduce correlation among bids. In an ascending auction, equilibrium bids satisfy $B_i = U_i$ regardless of whether bidders observe only their own valuations or also factors shifting all bidders' valuations. Hence it will be impossible to distinguish an environment with unobserved heterogeneity from an environment with correlated private values but no unobserved heterogeneity. In a first-price auction, the situation is similar. As long as the conditions of Theorem 5.2 hold, the data can be rationalized by equilibrium bidding. However, unobserved heterogeneity can account for some or all of the observed correlation (if any) among bids.

This observational equivalence can be important. If (6.10) holds, for example, an assumption about whether each bidder i observes only U_i or also V_0 can have significant implications for bidding strategies (and, therefore, the appropriate interpretation of bids) in a first-price auction. Thus, one must rely on an assumption regarding which model is appropriate.[47] In her application to highway procurement auctions, Krasnokutskaya (2004) compares the estimated bid function under an assumption of affiliated private values (with no unobserved heterogeneity) to the average (over the unobserved heterogeneity) bid function under the assumption of independent private values with unobserved heterogeneity. She finds that the estimated average bid function under unobserved heterogeneity is steeper than the estimated bid function under affiliated private values, and that estimated average markups are substantially higher when one ignores unobserved heterogeneity. Hence, the modeling choice can have important implications.

The second issue is whether the joint distribution $F_{\mathbf{U}}(\cdot)$, is identified under the assumption of unobserved heterogeneity. We will see that in a first-price sealed-bid auction, identification can be obtained through additional structure, e.g., on the statistical and functional relationships between the unobserved heterogeneity and bidder valuations or on the effects of unobserved heterogeneity on observed outcomes other than bids. In an ascending auction, the available identification results require additional sources of variation in the data, such as bidder-specific covariates.

The third issue is whether identification of $F_{\mathbf{U}}(\cdot)$ is adequate for the economic questions one wishes to answer. In the presence of unobserved heterogeneity, knowledge of this distribution is sufficient to answer some important questions but not others – in particular, not those concerning outcomes that depend on bidders' beliefs about opponents' valuations, since these beliefs vary with the realization of the factor that is unobservable to the econometrician. In ascending or second-price auctions (or any mechanism with a dominant strategy equilibrium), $F_{\mathbf{U}}(\cdot)$ is the only primitive relevant for predicting equilibrium outcomes, designing the auction rules, or performing counterfactual simulations. However, if we wish to consider policy questions concerning first-price auctions or other mechanisms in which beliefs play a more significant role, it will be necessary to know the joint distribution of bidders' private information and the unobserved heterogeneity (e.g., the distribution $F_{\mathbf{A},V_0}(\cdot)$ if one assumes (6.10)), not just $F_{\mathbf{U}}(\cdot)$. Below we will discuss conditions under which this joint distribution is identified.

[47] Although the literature has not yet considered approaches for distinguishing between the two models, it may be possible to develop a test based on exogenous variation in participation. In particular, it is possible to estimate the primitives of each model for a fixed set of potential bidders $\mathcal{N}$. Then, these primitives can be used to make "out of sample" predictions about bid distributions for other sets of potential bidders (e.g., a subset of the original set $\mathcal{N}' \subset \mathcal{N}$). We conjecture that in general the specific bid distributions predicted by the two models for the set of bidders $\mathcal{N}'$ will differ across the two models. However, to our knowledge this has not been formally analyzed. Note that a test of this restriction would rely on the assumption that participation does not vary with the unobserved heterogeneity. This assumption may be strong in practice; it may be satisfied, however, if bidders pay a cost to acquire a signal and the unobserved heterogeneity is not observed by bidders until they bear the cost of investigating the auction. Instrumental variables approaches like that explored in Haile, Hong and Shum (2003) may also be useful.

6.1.2.1. First-price sealed-bid auctions To demonstrate the problem of unobserved heterogeneity in first-price auctions, we begin with a very general case. In a private values first-price sealed-bid auction, suppose that information $\mathbf{w}_t$ is common knowledge among the bidders at auction t. Following the discussion in Section 6.1.1, the first-order condition relating bids to the underlying valuations is

$$u_{it} = b_{it} + \frac{G_{M_i|B_i}(b_{it}|b_{it}; \mathcal{N}, \mathbf{w}_t)}{g_{M_i|B_i}(b_{it}|b_{it}; \mathcal{N}, \mathbf{w}_t)}. \tag{6.11}$$

If the econometrician does not observe $\mathbf{w}_t$, the conditional distribution $G_{M_i|B_i}(b_{it}|b_{it}; \mathcal{N}, \mathbf{w}_t)$ is not identified. This creates a serious challenge to any attempt to uncover the markdown $\frac{G_{M_i|B_i}(b_{it}|b_{it};\mathcal{N},\mathbf{w}_t)}{g_{M_i|B_i}(b_{it}|b_{it};\mathcal{N},\mathbf{w}_t)}$. Indeed, because this markdown is a nonlinear function of $\mathbf{w}_t$, even the average markdown is not identified in general [Hendricks, Pinkse and Porter (2003)].

Identification requires additional structure, and several possibilities have been explored in the literature. All begin by assuming that the unobservable is a scalar, which we will denote by W. Some early work took parametric approaches to disentangling the common shock W from idiosyncratic factors, but more recently nonparametric identification results have been derived, exploiting additional data and/or assumptions about the way common shocks affect outcomes.

One approach, first proposed by Campo, Perrigne and Vuong (2003), is to exploit observables that are sufficient for the unobserved factor. This can be natural when there is an observable endogenous variable besides bids that responds to the unobservable W.[48] Both Campo, Perrigne and Vuong (2003) and Haile, Hong and Shum (2003) have used this approach by positing a model in which the number of bidders in auction t can be represented as a function of observables $\mathbf{Z}_t$ and the unobservable W_t:

$$N_t = \alpha(\mathbf{Z}_t, W_t). \tag{6.12}$$

If $\alpha(\mathbf{z}, \cdot)$ is a strictly increasing function for all $\mathbf{z}$, then the joint distribution of $(X_1, \ldots, X_n, U_1, \ldots, U_n)$ conditional on $(N_t, \mathbf{Z}_t)$ is identical to that conditional on $(N_t, \mathbf{Z}_t, W_t)$. Then

$$\begin{aligned} v(x, x; n, w, z) &\equiv E[U_i \mid X_i = x,\ N_t = n,\ W_t = w,\ \mathbf{Z}_t = \mathbf{z}] \\ &= E[U_i \mid X_i = x,\ N_t = n,\ \mathbf{Z}_t = \mathbf{z}] \\ &\equiv v(x, x; n, \mathbf{z}) \end{aligned}$$

and identification follows from the first-order condition

[48] A similar idea was used to address the problem of identification with unobserved heterogeneity in a very different environment by Olley and Pakes (1996).

$$v(x_{it}, x_{it}; n_t, \mathbf{z}_t) = b_{it} + \frac{\Pr(\max_{j\neq i} B_{jt} \leqslant b_{it} \mid B_{it} = b_{it},\ \mathbf{Z}_t = \mathbf{z}_t,\ N_t = n_t)}{\frac{\partial}{\partial m}\Pr(\max_{j\neq i} B_{jt} \leqslant m \mid B_{it} = b_{it},\ \mathbf{Z}_t = \mathbf{z}_t,\ N_t = n_t)|_{m=b_{it}}},$$

where the right-hand side is a known function of observables.

The assumption of strict monotonicity of N in W is strong although it is clear that there must be an invertible relation between W and the observables for this kind of approach. With weak monotonicity, conditioning on $(N_t, \mathbf{Z}_t)$ would limit the realization of W_t to some set $\mathcal{W}(N_t, \mathbf{Z}_t)$, and in some applications this might be sufficient to use the first-order condition above as a useful approximation.

The economic interpretation of (6.12) can be important when taking this kind of approach. For example, to predict outcomes under alternative selling mechanisms, one must consider whether changing mechanisms would alter the relation between bidder participation and $\mathbf{Z}$ [see, e.g., Athey, Levin and Seira (2004)]. If so, one would need a fully specified economic model of participation and bidding. However, a reduced form may be adequate for some questions and applications – for example, when (6.12) describes the determination of matches between auctions and potential bidders based on unobserved characteristics of the object offered for sale, or when the economic questions of interest do not depend on counterfactual predictions regarding participation.

Other approaches to handling unobserved heterogeneity in a first-price auction are closely related to ideas from the econometrics literatures on measurement error with repeated measures [Li and Vuong (1998), Li (2002), Schennach (2004)] and duration models with unobserved heterogeneity and multiple spells [see, e.g., Lancaster (1990)]. These literatures consider multiple observations for each of many units, with observations within each unit reflecting both a common (unobserved) shock as well as idiosyncratic shocks. In the auction setting, the auction plays the role of the unit, with the individual bids being the observations within unit.

Consider a simplified model of unobserved heterogeneity in which bidder valuations take the additively separable form in (6.10), and $(A_1, \ldots, A_n, V_0)$ are mutually independent with compact support. This is a special case of a *conditionally independent private values* model (itself a special case of affiliated private values, so long as each $f_{A_i}(\cdot)$ is log-concave).[49] Li, Perrigne and Vuong (2000) considered this structure under the assumption that bidders observe only their valuations U_i. They showed that, in that case, the joint distribution $F_{\mathbf{A},V_0}(\cdot)$ is nonparametrically identified up to a location normalization. While there is no unobserved heterogeneity in their model, their approach

[49] In a general specification of conditionally independent private values, one would assume only $F_{\mathbf{A},V_0}(a_1, \ldots, a_n, v_0) = F_{V_0}(v_0)\prod_{i=1}^n F_{A_i}(a_i|v_0)$. With this more general specification, the linearity assumed in (6.10) would be without loss of generality, since whatever the distributions of $U_i|V_0$, one can let $A_i = U_i - V_0$. The de Finetti Theorem [e.g., Chow and Teicher (1997)] tells us that any infinite sequence of exchangeable random variables can be represented by this more general conditionally independent structure. However, finite exchangeable sequences, like those arising in symmetric auctions with a finite number of potential bidders, need not have such a representation. Athey and Haile (2000, Proposition 4) explore limitations of the flexibility of the more restrictive conditionally independent structure considered here.

turns out to be a useful starting point. To see the idea behind their result, recall that observation of all bids and bidder identities is sufficient to identify the joint distribution $F_{\mathbf{U}}(\cdot)$ in a first-price auction with affiliated private values. Once $F_{\mathbf{U}}(\cdot)$ is known, a result from the literature on measurement error can be applied to separately identify the component distributions $F_{\mathbf{A}}(\cdot)$ and $F_{V_0}(\cdot)$ up to a location normalization. Li, Perrigne and Vuong (2000) develop consistent nonparametric estimators for this environment using empirical characteristic functions.[50]

Krasnokutskaya (2004) shows that a very similar approach can be applied in the case of unobserved heterogeneity – i.e., when valuations take the additively separable form in (6.10) and v_0 is observed by bidders but not the econometrician.[51] In essence, she reverses the two steps of Li, Perrigne, and Vuong's (2000) approach: she first uses a deconvolution technique to remove the effects of unobserved heterogeneity from bids, then recovers the idiosyncratic factors a_i through the first-order condition for a hypothetical auction with no unobserved heterogeneity. In this sense, the approach is similar to the "homogenization" approach for incorporating observable auction heterogeneity, discussed in Section 6.1.1.

For the first step, recall from Section 6.1.1 that the additive separability in (6.10) is preserved by equilibrium bidding.[52] So if $\beta_i(u_{it}; \mathcal{N}_t, v_{0t})$ denotes bidder i's equilibrium bid given $u_{it}, \mathcal{N}_t$, and v_{0t}, then

$$\beta_i(u_{it}; \mathcal{N}_t, v_{0t}) = \beta_i(u_{it} - v_{0t}; \mathcal{N}_t, 0) + v_{0t}. \tag{6.13}$$

If one observes all bids from each auction, the following result from Kotlarski (1966) implies identification of the joint distribution of $(\beta_1(A_1; \mathcal{N}, 0), \ldots, \beta_n(A_n; \mathcal{N}, 0), V_0)$ up to a location normalization.[53]

LEMMA 6.1. *Let Y_1, Y_2, and Y_3 be mutually independent random variables with nonvanishing characteristic functions $\phi_1(\cdot)$, $\phi_2(\cdot)$, and $\phi_3(\cdot)$, respectively. Let $Q_1 = Y_1 + Y_3$, $Q_2 = Y_2 + Y_3$. Then*

(i) *the joint distribution of (Q_1, Q_2) completely determines the distributions of Y_1, Y_2, and Y_3 up to location;*

(ii) *if $\psi(\cdot,\cdot)$ denotes the characteristic function of (Q_1, Q_2), and $\psi_i(\cdot,\cdot)$ is its derivative with respect to its ith argument, then under the normalization $E[Y_1] = 0$, $\phi_3(t) = \exp\{\int_0^t \frac{\psi_1(0,s)}{\psi(0,s)}\, ds\}$, $\phi_1(t) = \frac{\psi(t,0)}{\phi_3(t)}$, and $\phi_2(t) = \frac{\psi(0,t)}{\phi_3(t)}$.*

[50] See also the discussion of a closely related special case of the mineral rights model in Section 7.2.1 below.

[51] An alternative is discussed in Section 8.2 below.

[52] Krasnokutskaya (2004) focuses on the case of multiplicative rather than additive separability. The analysis is equivalent with a logarithmic transformation. As she points out, a more general model allowing unobserved heterogeneity affecting both the location and scale of private values is also identifiable, since one can apply the deconvolution step (Lemma 6.1) to the bids twice – once in logs and once in levels.

[53] To see the connection to the original measurement error framework, observe that with an appropriate location normalization, under (6.10) each A_i can be interpreted as an independent mean-zero measurement error on v_0.

A proof can be found in Prakasa-Rao (1992, Theorem 2.1.1 and Remark 2.1.11).[54] A key to the result is the fact that the characteristic function of the sum of independent random variables is the product of the characteristic functions of the component variables. With multiple observations involving one component in common, this separability can be exploited to isolate the characteristic functions (and, thereby, the distributions) of the individual components. Identification is up to a location normalization, since adding a constant to Y_3 and subtracting the same constant from Y_1 and Y_2 has no effect on observables.

Lemma 6.1 can be used to relate characteristic functions of the observed bids to those of the "homogenized" bids $\beta_1(A_1;\mathcal{N},0),\ldots,\beta_n(A_n;\mathcal{N},0)$ and the unobserved factor V_0. Identification of the distribution of each A_i then follows from the first-order condition for a hypothetical auction in which $v_{0t}=0$. In particular, if we let $B_i^0 = \beta_i(A_i;\mathcal{N},0) = B_i - V_0$,

$$A_i = B_i^0 + \frac{\Pr(\max_{j\neq i}\beta_j(A_j;\mathcal{N},0)\leqslant B_i^0)}{\frac{\partial}{\partial m}\Pr(\max_{j\neq i}\beta_j(A_j;\mathcal{N},0)\leqslant m)|_{m=B_i^0}} \equiv \tilde{\xi}_i\big(B_i^0;\mathcal{N}\big). \tag{6.14}$$

Note that unlike the case without unobserved heterogeneity, it is not possible to identify the valuations of bidders in a particular auction, because the realization of V_0 is unobserved. Despite this, because Lemma 6.1 implies that $G_{B_i^0}(\cdot)$ is identified, it follows that $\tilde{\xi}_i(\cdot;\mathcal{N})$ is also identified, so that the distribution of private information is given by

$$F_{A_i}(a_i) = G_{B_i^0}\big(\tilde{\xi}_i^{-1}(a_i;\mathcal{N})\big).$$

Nonparametric estimators can be developed by first substituting empirical characteristic functions for the population characteristic functions in part (ii) of Lemma 6.1, and then using simulation to construct pseudo-draws of the random variable on the right-hand-side of (6.14). We sketch the approach here. For simplicity, consider the special case in which there are two classes of bidders, with bidders in the same class drawing their valuations from the same marginal distribution (extension to more than two types is straightforward). Suppose one has a sample of T auctions in which there are n_1 class-1 and n_2 class-2 bidders in each auction, and that all bids and bidder identities are observable. As above, estimation must be undertaken fixing the number of bidders of each type, which is equivalent here to fixing the set $\mathcal{N}$.

Let $c(j,t)$ denote the class of bidder j in auction t. Impose the normalization $E[A_i]=0$ for any class-1 bidder i. Let $G_{B^j}(\cdot)$ denote the marginal distribution of the equilibrium bid B^j of a class-j bidder, and let $G_{B^1,B^2}(b^1,b^2)$ denote the joint distribution of (B^1,B^2). Similarly, let $B^{0,j}\equiv B^j - V_0$ denote the homogenized bid of a class-j bidder. Note that the homogenized bids are independent. Let $\psi(\cdot,\cdot)$, $\phi_0(\cdot)$,

[54] See also Li and Vuong (1998, Lemma 2.1).

$\phi_{B^{0,1}}(\cdot)$ and $\phi_{B^{0,2}}(\cdot)$ denote the characteristic functions of (B^1, B^2), V_0, $B^{0,1}$, and $B^{0,2}$, respectively.

Following Li and Vuong (1998) and Krasnokutskaya (2004) [see also Li, Perrigne and Vuong (2000)], define estimators

$$\hat{\psi}(\tau_1, \tau_2) = \frac{1}{Tn_1n_2} \sum_{t=1}^{T} \sum_{j:\, c(j,t)=1} \sum_{k:\, c(k,t)=2} \exp(\mathrm{i}\tau_1 b_{jt} + \mathrm{i}\tau_2 b_{kt}),$$

$$\hat{\psi}_1(\tau_1, \tau_2) = \frac{1}{Tn_1n_2} \sum_{t=1}^{T} \sum_{j:\, c(j,t)=1} \sum_{k:\, c(k,t)=2} \mathrm{i}b_{jt} \exp(\mathrm{i}\tau_1 b_{jt} + \mathrm{i}\tau_2 b_{kt}),$$

where, for each estimator, an average is taken over all possible pairs (b^1, b^2). Let

$$\hat{\phi}_0(\tau) = \exp\left\{ \int_0^{\tau} \frac{\hat{\psi}_1(0, v)}{\hat{\psi}(0, v)} \,\mathrm{d}v \right\},$$

$$\hat{\phi}_{B^{0,1}}(\tau) = \frac{\hat{\psi}(\tau, 0)}{\hat{\phi}_0(\tau)},$$

$$\hat{\phi}_{B^{0,2}}(\tau) = \frac{\hat{\psi}(0, \tau)}{\hat{\phi}_0(\tau)}.$$

Given these estimated characteristic functions, one can obtain estimates of the marginal densities of $B^{0,1}$, $B^{0,2}$ and V_0 using the inverse Fourier transform. In particular, let

$$\hat{g}_{B^{0,i}}(b) = \frac{1}{2\pi} \int_{-\mu}^{\mu} \exp(-\mathrm{i}\tau b) \hat{\phi}_{B^{0,i}}(\tau) \,\mathrm{d}\tau, \tag{6.15}$$

$$\hat{f}_{V_0}(v) = \frac{1}{2\pi} \int_{-\mu}^{\mu} \exp(-\mathrm{i}\tau v) \hat{\phi}_{V_o}(\tau) \,\mathrm{d}\tau, \tag{6.16}$$

where μ is a trimming parameter.

As shown by Li and Vuong (1998), under certain smoothness conditions (6.15) and (6.16) provide uniformly consistent estimators of the density $f_{V_0}(\cdot)$ of V_0 and the densities of the homogenized bids B_i^0 for each bidder i. These densities can then be used to construct estimates of the right-hand-side of the first-order condition (rewriting (6.14))

$$A_{it} = B_{it}^0 + \frac{\prod_{j \neq i} G_{B^{0,c(j,t)}}(B_{it}^0)}{\sum_{j \neq i} g_{B^{0,c(j,t)}}(B_{it}^0) \prod_{k \neq i,\, j} G_{B^{0,c(k,t)}}(B_{it}^0)}, \tag{6.17}$$

where

$$G_{B^{0,j}}(b) = \int_{-\infty}^{b} g_{B^{0,j}}(s) \,\mathrm{d}s. \tag{6.18}$$

In contrast to other applications of the indirect approach to first-price auctions [e.g., Guerre, Perrigne and Vuong (2000)], however, here draws of the bids B_{it}^0 on the right-hand-side of (6.17) cannot be taken directly from the data. Instead, they must be

simulated from the estimated densities $\hat{g}_{B_i^0}(b)$. Using simulated bids, (6.17) makes it possible to construct a pseudo-sample of draws of the idiosyncratic components A_i, which can be used to obtain estimates of their underlying densities $f_{A^j}(\cdot)$ using standard methods. Krasnokutskaya (2004) provides additional details and conditions under which this leads to uniformly consistent estimates of the marginal densities $f_{V_0}(\cdot)$ and $f_{A^j}(\cdot)$ for each bidder class j. She suggests the use of the bootstrap for inference.

Note that while the approach here is similar to that in Li, Perrigne and Vuong (2000), there are important distinctions. When V_0 is not observed by bidders, the joint distribution $F_{\mathbf{U}}(\cdot)$ is identified directly from the first-order condition and completely characterizes bidder demand and information. Since knowledge of $F_{\mathbf{U}}(\cdot)$ is sufficient for counterfactual simulations in a private values model with no unobserved heterogeneity, it is not clear under what circumstances one would need to separately identify $F_{\mathbf{A}}(\cdot)$ and $F_{V_0}(\cdot)$.[55] When V_0 is observed by the bidders, however, identification of the joint distribution $F_{\mathbf{U}}(\cdot)$ no longer follows directly from the first-order condition. Furthermore, even if $F_{\mathbf{U}}(\cdot)$ were identified, in this environment separate identification of $F_{\mathbf{A}}(\cdot)$ and $F_{V_0}(\cdot)$ is required for many counterfactuals.

The approach proposed by Krasnokutskaya (2004) is attractive in that it places no restriction on the distribution of the idiosyncratic factor A_i or the distribution of V_0. It does restrict the way unobservables affect valuations. It may also require large samples – the slow convergence rates of deconvolution estimators is well known. Athey, Levin and Seira (2004) propose an alternative, trading flexibility in the specifications of $F_{V_0}(\cdot)$ and the $F_{A_i}(\cdot)$ for flexibility in how unobservable and observable auction characteristics affect valuations. They propose parametric estimation of the bid distributions and the distribution of auction heterogeneity. This is followed by estimation of the distribution of valuations based on (6.14) in a manner similar to Krasnokutskaya (2004). Mixtures of parametric models might be introduced to allow for more flexibility, as described at the end of Section 6.1.1. Although using a parametric first step is restrictive, it allows a parsimonious specification whereby the unobserved heterogeneity may affect some types of bidders differently than others, and where the distribution of the unobserved heterogeneity depends on auction characteristics. In principle, these features could be incorporated into Krasnokutskaya's (2004) approach by allowing auction characteristics to interact with V_0 and A_i in (6.10), but in practice this may be not be feasible in data sets of moderate size.

6.1.2.2. Ascending auctions The challenges created by unobserved auction heterogeneity in an ascending auction are quite different. Because equilibrium is in weakly dominant strategies in the standard model of the ascending auction, unobserved heterogeneity does not affect the equilibrium mapping (the identity function) between

[55] A separate (and open) question, however, is whether imposing the structure of this model in estimation leads to more precise estimates in counterfactual simulations.

valuations and bids. For example, bidding in an environment with valuations characterized by (6.10) is the same regardless of whether bidders observe both v_0 and a_i or only their sum. The main problem posed by such an environment is the fact that positive identification results for ascending auctions have been obtained primarily for environments with independent valuations, yet the presence of an unobserved factor like v_0 generally leads to a violation of independence.

In Section 6.2.1 we will show how additional data on bidder characteristics can be used to obtain identification of the joint distribution of valuations in an ascending auction without independence. This would not be sufficient for all economic questions of interest, however. As the preceding section makes clear, for example, separate identification of $F_{V_0}(\cdot)$ and each $F_{A_i}(\cdot)$ is needed even to simulate outcomes in a first-price sealed-bid auction. However, with an estimate of the joint distribution $F_{\mathbf{U}}(\cdot)$, it should be possible to use deconvolution techniques similar to those discussed above to separately estimate $F_{V_0}(\cdot)$ and each $F_{A_i}(\cdot)$ when $U_i = A_i + V_0$, under assumptions similar to those discussed above. This has not yet been investigated.

6.2. Bidder heterogeneity

6.2.1. Observed bidder heterogeneity

As discussed in prior sections, observable differences across bidders introduce asymmetry that can complicate the analysis of bidding data. However, when bidder-specific covariates are observable and vary across auctions, they can actually aid identification by enabling the distribution function for a single order statistic to reveal more information. This is particularly valuable in an ascending auction given the negative identification results above for environments without independence. In fact, with sufficiently rich variation in covariates, identification can be obtained with asymmetric dependent valuations, even when the transaction price is the only bid available (or the only bid assumed to have the unambiguous interpretation implied by the button auction model).

The idea behind this approach is familiar from other types of models, including the Roy model of labor supply [e.g., Heckman and Honoré (1990)] and competing risks models [e.g., Heckman and Honoré (1989)]. To see how this can work in the auction environment, suppose

$$U_i = g_i(W_i) + A_i,$$

where each $g_i(\cdot)$ is an unknown function, W_i is a covariate reflecting characteristics of bidder i, and the private stochastic components $(A_1, \ldots, A_n)$ are drawn from an arbitrary joint distribution $F_{\mathbf{A}}(\cdot)$ and are independent of the matrix $\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$. Suppose for the moment that each $g_i(\cdot)$ is known and that we could somehow observe

$u^{(n:n)}$.[56] Conditional on the vector $\mathbf{w}$, $U^{(n:n)}$ has cumulative distribution

$$\begin{aligned} F_U^{(n:n)}(u|\mathbf{w}) &= \Pr\bigl(U^{(n:n)} \leqslant u \mid \mathbf{w}\bigr) \\ &= F_{\mathbf{U}}(u, \ldots, u|\mathbf{w}) \\ &= \Pr\bigl(g_i(w_i) + A_i \leqslant u \ \forall i\bigr) \\ &= F_{\mathbf{A}}\bigl(u - g_1(\mathbf{w}_1), \ldots, u - g_n(\mathbf{w}_n)\bigr). \end{aligned}$$

While the joint distribution $F_{\mathbf{U}}(\cdot|\mathbf{w})$ is observed only along the diagonal ($U_1 = \cdots = U_n$), sufficient variation in $(g_1(\mathbf{w}_1), \ldots, g_n(\mathbf{w}_n))$ would "trace out" the entire joint distribution $F_{\mathbf{A}}(\cdot)$. Furthermore, prior knowledge of the functions $g_i(\cdot)$ is not necessary with sufficient variation in covariates: at sufficiently large negative values of $g_j(w_j)$ $\forall j \neq i$, bidder i will have the largest valuation with probability arbitrarily close to one, so that variation in $\mathbf{w}_i$ and the point of evaluation u would trace out the function $g_i(\cdot)$.

In practice we cannot observe $u^{(n:n)}$ in an ascending auction, and the distribution of an interior order statistic has a more complicated relation to the underlying joint distribution than does the maximum (or minimum, as in the case of competing risks). However, the following result shows that the fundamental idea behind this approach can be used to obtain identification in an ascending auction when only the transaction price is observable.[57]

THEOREM 6.1. *Assume*

(i) $U_i = g_i(\mathbf{W}_i) + A_i$, $i = 1, \ldots, n$.
(ii) $F_{\mathbf{A}}(\cdot)$ *has support* $\mathbb{R}^n$ *and a continuously differentiable density.*
(iii) A_i *and* $\mathbf{W}_j$ *are independent for all* i, j.
(iv) $\operatorname{supp}(g_1(\mathbf{W}_1), \ldots, g_n(\mathbf{W}_n)) = \mathbb{R}^n$.
(v) *For all* i, $g_i(\cdot)$ *is continuously differentiable, with* $\lim_{\mathbf{w}_i \to (\infty,\ldots,\infty)} g_i(\mathbf{w}_i) = \infty$ *and* $\lim_{\mathbf{w}_i \to (-\infty,\ldots,-\infty)} g_i(\mathbf{w}_i) = -\infty$.

Then $F_{\mathbf{A}}(\cdot)$ *and each* $g_i(\cdot)$, $i = 1, \ldots, n$, *are identified up to a location normalization from observation of* $U^{(j:n)}$ *and* $\mathbf{W}$, *for any single value of* $j \in \{1, \ldots, n\}$.

PROOF. For simplicity let each $\mathbf{W}_i = W_i$ be a scalar. For $\mathcal{T} \subset \{1, \ldots, n\}$ define

$$\overline{F}_{\mathbf{A}}^{\mathcal{T}}(a_1, \ldots, a_n) \equiv \Pr(A_i > a_i \ \forall i \in \mathcal{T},\ A_j \leqslant a_j \ \forall j \notin \mathcal{T})$$

[56] This is the observable order statistic in the Roy model, where the wage in the chosen sector (the one offering the highest wage) is the only one observed, yet one is interested in the joint distribution of wage offers from all sectors.

[57] The result is a slight modification of Theorem 5 of Athey and Haile (2002), correcting a minor error in their proof.

and let $\overline{F}^{\mathcal{T}}_{\mathbf{A},A_i}(a_1,\dots,a_n) = \frac{\partial}{\partial a_i}\overline{F}^{\mathcal{T}}_{\mathbf{A}}(a_1,\dots,a_n)$. For arbitrary $u \in \mathbb{R}$, define $\mathbf{z} = (u - g_1(w_1),\dots,u - g_n(w_n))$. Then

$$\Pr\bigl(U^{(j:n)} \leqslant u \mid \mathbf{w}\bigr) = \sum_{\substack{\mathcal{T}\subset\{1,\dots,n\}\\ |\mathcal{T}|=n-j}} \sum_{i\notin\mathcal{T}} \int_{-\infty}^{u} \overline{F}^{\mathcal{T}}_{\mathbf{A},A_i}\bigl(\tilde{u} - g_1(w_1),\dots,\tilde{u} - g_n(w_n)\bigr)\,\mathrm{d}\tilde{u},$$

where the summations are over the possible identities of the bidders with the $n - j$ highest bids, and the identity of the bidder i with bid $B^{(j:n)}$. Differentiation yields

$$\begin{aligned}
&\frac{\partial}{\partial u}\frac{\partial^n}{\partial w_1\cdots\partial w_n}\Pr\bigl(U^{(j:n)} \leqslant u \mid \mathbf{w}\bigr)\\
&\quad= \sum_{\substack{\mathcal{T}\subset\{1,\dots,n\}\\ |\mathcal{T}|=n-j}} \sum_{i\notin\mathcal{T}} (-1)^{n-j} \prod_{k=1}^{n}\bigl(-g_k'(w_k)\bigr)\frac{\partial}{\partial a_i} f_{\mathbf{A}}(\mathbf{a})\Big|_{\mathbf{a}=\mathbf{z}}\\
&\quad= \binom{n-1}{n-j}(-1)^{n-j}\prod_{k=1}^{n}\bigl(-g_k'(w_k)\bigr)\sum_{i=1}^{n}\frac{\partial}{\partial a_i} f_{\mathbf{A}}(\mathbf{a})\Big|_{\mathbf{a}=\mathbf{z}}
\end{aligned}$$

since there are $\binom{n-1}{n-j}$ subsets $\mathcal{T}$ of size $n - j$ that exclude i. Now observe that

$$\begin{aligned}
&\frac{\partial}{\partial u}\frac{\partial^n}{\partial w_1\cdots\partial w_n} F_{\mathbf{A}}\bigl(u - g_1(w_1),\dots,u - g_n(w_n)\bigr)\\
&\quad= \prod_{k=1}^{n}\bigl(-g_k'(w_k)\bigr)\sum_{i=1}^{n}\frac{\partial}{\partial a_i} f_{\mathbf{A}}(\mathbf{a})\Big|_{\mathbf{a}=\mathbf{z}}\\
&\quad= \frac{1}{\binom{n-1}{n-j}(-1)^{n-j}}\frac{\partial}{\partial u}\frac{\partial^n}{\partial w_1\cdots\partial w_n}\Pr\bigl(U^{(j:n)} \leqslant u \mid \mathbf{w}\bigr).
\end{aligned}$$

Hence, using the fundamental theorem of calculus,

$$\begin{aligned}
&\frac{\partial^n}{\partial w_1\cdots\partial w_n} F_{\mathbf{A}}\bigl(u - g_1(w_1),\dots,u - g_n(w_n)\bigr)\\
&\quad= \frac{1}{\binom{n-1}{n-j}(-1)^{n-j}}\frac{\partial^n}{\partial w_1\cdots\partial w_n}\Pr\bigl(U^{(j:n)} \leqslant u \mid \mathbf{w}\bigr).
\end{aligned}$$

Repeated application of the fundamental theorem of calculus shows that

$$\begin{aligned}
&\int_{w_1}^{\infty}\cdots\int_{w_n}^{\infty}\frac{\partial^n}{\partial\tilde{w}_1\cdots\partial\tilde{w}_n} F_{\mathbf{A}}\bigl(u - g_1(\tilde{w}_1),\dots,u - g_n(\tilde{w}_n)\bigr)\,\mathrm{d}\tilde{w}_n\ldots\mathrm{d}\tilde{w}_1\\
&\quad= (-1)^n F_{\mathbf{A}}\bigl(u - g_1(w_1),\dots,u - g_n(w_n)\bigr)
\end{aligned}$$

so that

$$F_{\mathbf{A}}\big(u - g_1(w_1), \ldots, u - g_n(w_n)\big)$$
$$= \frac{(-1)^j}{\binom{n-1}{n-j}} \int_{w_1}^{\infty} \cdots \int_{w_n}^{\infty} \frac{\partial^n}{\partial \tilde{w}_1 \cdots \partial \tilde{w}_n} \Pr\big(U^{(j:n)} \leqslant u \mid \tilde{\mathbf{w}}\big)\, \mathrm{d}\tilde{w}_n \ldots \mathrm{d}\tilde{w}_1. \tag{6.19}$$

Now note that $\lim_{\mathbf{w}_{-i} \to (-\infty,\ldots,-\infty)} F_{\mathbf{A}}(u - g_1(w_1), \ldots, u - g_n(w_n)) = F_{A_i}(u - g_i(w_i))$, where $F_{A_i}(\cdot)$ is the marginal distribution of A_i. For each i, then, variation in u and w_i identifies $g_i(\cdot)$ through Equation (6.19) by standard arguments. With knowledge of each $g_i(\cdot)$ we can then use (6.19) to uniquely determine $F_{\mathbf{A}}(\cdot)$ at any point $(a_1, \ldots, a_n)$ through appropriate choices of u and $\mathbf{w}$. □

Estimation based on this result has not yet been explored. For the competing risks model, however, Fermanian (2003) has recently proposed kernel methods that build directly on the closely related identification proof of Heckman and Honoré (1989).

6.2.2. *Unobserved bidder heterogeneity*

We have already discussed several models with bidder heterogeneity that is either fixed across all auctions or captured by observable bidder-specific covariates. However, one can imagine situations in which asymmetries between bidders vary across auctions due to factors that are common knowledge to bidders but unobserved to the econometrician. For example, the match between the specifications of a procurement contract and each contractor's particular expertise might be common knowledge within the industry but unobservable to outsiders.

In the most general case, this type of environment requires a different marginal distribution $F_{U_{it}}(\cdot)$ for each bidder i's valuation in each auction t. It should be clear that identification of such a model from bid data alone is impossible: the number of marginal distributions in the model is equal to the number of observations, even assuming one observes all bids from each auction.

Consider instead a more restrictive model

$$U_{it} = A_{it} + E_{it},$$

where all (i) A_{it} are i.i.d. draws from a cumulative distribution $F_A(\cdot)$ with density $f_A(\cdot)$; (ii) E_{it} and A_{it} are mutually independent; (iii) E_{it} is common knowledge among the bidders but unobserved to the econometrician; and (iv) each E_{it} is an independent draw from a cumulative distribution $F_{E_i}(\cdot)$ with density $f_{E_i}(\cdot)$. From the econometrician's perspective, each bidder's valuation is then an independent draw from a density

$$f_{U_i}(\cdot) = f_A(\cdot) * f_{E_i}(\cdot),$$

where $*$ denotes convolution.

In an ascending auction, Theorem 4.1 implies that each $F_{U_i}(\cdot)$ is identified if one observes the transaction price, the set of bidders $\mathcal{N}$, and the winner's identity. This would be sufficient for some important questions and policy simulations, although not all. For

example, it would not be sufficient to simulate outcomes under a first-price sealed-bid auction, since to do this one would need to know how much of the variation in valuations was common knowledge (through E_i) and how much was private information (through A_i). Separate identification of $F_A(\cdot)$ and $F_{E_i}(\cdot)$ for all i is not possible from bid data, however. There are $n+1$ marginal distribution functions of interest. Yet even if one observed bids from all bidders (instead of the $n-1$ losing bids, as usually assumed), there are only n marginal distributions of observable bids. Without additional restrictions, identification will not be possible.

In a first-price auction, the situation is further complicated by the nontrivial strategic behavior. In particular, even identification of each $F_{U_i}(\cdot)$ in the special case above is doubtful, since the markdown in each bidder's first-order condition

$$u_i = b_i + \frac{\Pr(\max_{j\neq i} B_j \leqslant b_i \mid B_i = b_i,\ E_1, \ldots, E_j)}{\frac{\partial}{\partial m}\Pr(\max_{j\neq i} B_j \leqslant b_i \mid B_i = b_i,\ E_1, \ldots, E_j)|_{m=b_i}}$$

involves expectations that are conditioned on the information $E_1, \ldots, E_j$ that is unobservable to the econometrician.[58] The problem here is closely related to that discussed in Section 6.1.2, although the dimensionality of the unobserved heterogeneity is higher, and the approaches thus far proposed to address unobserved heterogeneity do not appear to be applicable.

6.3. Endogenous participation

So far, we have focused on models in which any variation in the set of bidders is exogenous (the exception is the discussion of endogenous participation with unobserved heterogeneity in Section 6.1.2). In this section we consider several different models of how the set of bidders is determined, and we explore the consequences of these models for identification. Here it will be useful to draw a distinction between *potential* bidders and *actual* bidders. As before, we let $\mathcal{N}$ (with $|\mathcal{N}| = n$) denote the set of potential bidders – those who draw signals and decide whether to bid.[59] We let $\mathcal{A} \subseteq \mathcal{N}$ (with $|\mathcal{A}| = a$) denote the set of actual bidders, i.e., those who actually place a bid. Variation in both $\mathcal{N}$ and $\mathcal{A}$ is possible. Let $\widetilde{\mathcal{N}}$ be the random set whose realization is denoted by $\mathcal{N}$, and let $\widetilde{\mathcal{A}}$ be the random set whose realization is denoted by $\mathcal{A}$.

[58] Models similar to this have been explored in the related context of differentiated products oligopoly price competition [e.g., Berry, Levinsohn and Pakes (1995); see also Chapter 63 by Ackerberg et al. in this volume]. There, common knowledge differences in unobservable (to the econometrician) quality of products that differ across markets lead to asymmetries in the effective common knowledge marginal costs of supplying utility to a buyer choosing between firms. Identification in those models is obtained through a combination of parametric assumptions and restrictions from the demand side of the market. In the auction setting, the latter would be analogous to restrictions from the seller's (or auctioneer's) side of the market, for example using the assumption that the reserve price is set optimally. We are not aware of empirical approaches exploiting such information, although this is a direction worth exploring. See Einav (2004) for a related discussion.

[59] In the literature, sometimes agents with the option of acquiring a signal are referred to as potential bidders [e.g., Hendricks, Pinkse and Porter (2003)].

An example of why the set of potential bidders may vary is an environment in which obtaining a signal is costly. Firms may then decide whether to investigate a particular opportunity at random or based on some summary statistics about the auction (for example, the appraised value of the object). Fixing the set of potential bidders, the set of actual bidders may vary, for example, if there is a binding reserve price or if submitting a bid is costly. In such cases, typically only bidders with sufficiently favorable signals will bid. In addition, in an ascending auction that lacks a strict "activity rule" like that in the standard Milgrom–Weber model, the set of actual bidders can exclude even potential bidders with relatively high valuations, since others may push the price beyond these bidders' willingness to pay before they ever make a bid.[60]

In this section we will see that the consequences of endogenous variation in $\mathcal{A}$ and $\mathcal{N}$ for equilibrium and identification will depend on whether bidders' participation decisions are common knowledge among the bidders and whether these are observable by the econometrician. Often the number of actual bidders in an auction is observed by the econometrician; the set of potential bidders may or may not be observed. [61]

6.3.1. Binding reserve prices

We first consider the case in which a reserve price may be binding. Recalling (2.1), in an n-bidder auction with reserve price r, only bidders with signals $x_i \geqslant x_i^*(r, \mathcal{N})$ participate (with $x_i^*(r, \mathcal{N}) = r$ in a private values auction). Ignoring this endogenous participation can result in misleading estimates due to the selection introduced by the participation decisions.[62] Throughout this section, we will assume $\mathcal{N}$ is observable, hold $\mathcal{N}$ fixed, and consider only bidders $i \in \mathcal{N}$.

6.3.1.1. Ascending auctions For ascending auctions we obtained positive identification results above primarily for models with independent private values (the exception is Theorem 6.1), so we will focus on such models here. Donald and Paarsch (1996)

[60] Auction-specific unobservables may affect either the number of potential bidders (e.g., if unobservables determine whether there is a suitable match between a specialized contractor and a contract offered by auction), or the number of actual bidders (e.g., if unobservables affect the profitability of an auction in an environment with costly signal acquisition). See Section 6.1.2 as well as Athey, Levin and Seira (2004), and Li and Zheng (2005).

[61] In the case that $\mathcal{N}$ is not observed but fixed in a sample, in most models of endogenous participation the common support assumption ensures that the union of identities of all actual bidders ever observed will converge to $\mathcal{N}$ as the sample of auctions grows [cf. Guerre, Perrigne and Vuong (2000)].

[62] A closely related model is that in which bidders must pay a fee to enter the auction [Samuelson (1985)] or, equivalently from the perspective of identification, preparing a bid is costly. This can lead to a participation rule very similar to that with a binding reserve price [Milgrom and Weber (1982)]. For first-price sealed-bid auctions, Haile, Hong and Shum (2003) discuss this case and provide results similar to those given in this section. Note that bid preparation costs are different from costs of acquiring a signal (discussed in Section 6.3.2), because in the former case a bidder places a bid if his signal is high enough, while in the latter case the participation decision must be made before bidders have obtained signals, and all bidders who acquire signals will bid (unless there is a binding reserve price).

and Paarsch (1997) were the first to incorporate reserve prices in structural models of ascending auctions in the IPV setting.[63] They observed that in a parametric framework one may account for the endogeneity of participation in one of two ways. First, if the number of potential bidders is observable, one may explicitly account (e.g., in a likelihood function) for the fact that the valuations (bids) of $(n - a)$ potential bidders were censored because these were below r. Alternatively, one can examine the bidding behavior of the actual bidders conditional on their decision to participate. This second approach is based on the fact that under independence each participating bidder i has a valuation that is an independent draw from the distribution

$$F_{U_i}(u|r) = \frac{F_{U_i}(u) - F_{U_i}(r)}{1 - F_{U_i}(r)}. \tag{6.20}$$

This observation is useful for considering nonparametric identification as well. With this observation, Theorem 4.1 implies that each truncated distribution $F_{U_i}(\cdot|r)$ is nonparametrically identified.

COROLLARY 6.1. *In an ascending auction with symmetric independent private values, $F_U(\cdot|r)$ is identified when the transaction price and the number of actual bidders is observable. In the asymmetric independent private values model, for each $i \in \mathcal{N}$, $F_{U_i}(\cdot|r)$ is identified when the transaction price, the identity of the winning bidder, and the set $\mathcal{A}$ are observable.*

In many cases, this result alone will be sufficient to enable one to address interesting questions. In the symmetric case, for example, Haile and Tamer (2003) have shown that the truncated distribution $F_U(\cdot|r)$ can be sufficient to determine the optimal reserve price (recall Equation (4.12)). To state the result, let $F_{U|r}(\cdot)$ denote $F_U(\cdot|r)$, and let c_0 be the value the seller places on the good (or her marginal cost of providing it).

THEOREM 6.2. *Given any univariate CDF $\Phi(\cdot)$, let $\pi(r; \Phi) = (r - c_0)(1 - \Phi(r))$ and $p^*(\Phi) \in \arg\max_{p \in \text{supp}\,\Phi(\cdot)} \pi(p; \Phi)$. Suppose $\pi(\cdot; F_U)$ is continuously differentiable and strictly quasi-concave. Then* (i) *if $r < p^*(F_U)$, $r^*(F_{U|r}) = p^*(F_U)$*; (ii) *if $r \geqslant p^*(F_U)$, $p^*(F_{U|r}) = r$.*

This result implies that in a symmetric IPV environment, the optimal reserve one would calculate by ignoring the endogenous participation is actually optimal, except when the actual reserve price results in truncation of the relevant region of support. This follows from the fact that the objective functions $\pi(\cdot; F_U)$ and $\pi(\cdot; F_{U|r})$ differ only by a multiplicative constant. The qualification concerning truncation is important but not surprising: if there are no data below the true optimal reserve price, this optimum cannot

[63] More recently, Donald, Paarsch and Robert (2006), and Bajari and Hortaçsu (2003a) have considered parametric models incorporating endogenous participation with reserve prices.

be detected. However, part (ii) of Theorem 6.2 ensures that when such truncation has occurred, the data will at least reveal this fact.

For some policy questions, including predicting revenues under a different mechanism or reserve price, the full (untruncated) distributions $F_{U_i}(\cdot)$ will be needed, even under the independent private values assumption. It should be clear that the value of $F_{U_i}(u)$ for u lower than all observed reserve prices could not be determined except through a parametric assumption. However, if both $\mathcal{N}$ and $\mathcal{A}$ are observable, each $F_{U_i}(u)$ can be recovered for all $u \geqslant r$. In particular, since $F_{U_i}(r) = \Pr(i \notin \widetilde{\mathcal{A}})$, identification of $F_{U_i}(u)$ for all $u \geqslant r$ follows immediately from (6.20) and Corollary 6.1.

THEOREM 6.3. *In the symmetric independent private values model, $F_U(u)$ is identified for all $u \geqslant r$ when the transaction price and $|\widetilde{\mathcal{A}}|$ are observable. In the asymmetric independent private values model, each $F_{U_i}(\cdot)$ is identified when the transaction price, the identity of the winning bidder, and $\widetilde{\mathcal{A}}$ is observable.*

An estimate of $F_{U_i}(u)$ for $u \geqslant r$ will be sufficient for some policy questions, e.g., calculations of revenues with higher reserve prices or under some alternative mechanisms. Estimation of each $F_{U_i}(r) = \Pr(i \notin \widetilde{\mathcal{A}})$ based on a sample analog is straightforward. In the case of symmetry, a different approach to estimation of $F_U(\cdot)$ is available: observe that exchangeability implies [Haile, Hong and Shum (2003)]

$$\begin{aligned} F_U(r) &= \Pr(U_1 \leqslant r) \\ &= F_{\mathbf{U}}(r, \infty, \ldots, \infty; n) \\ &= \sum_{k=1}^{n} \frac{k}{n} \Pr\bigl(|\widetilde{\mathcal{A}}| = n - k\bigr). \end{aligned} \tag{6.21}$$

A sample analog of (6.21) places much weaker demands on the data than a sample analog of $\Pr(i \notin \widetilde{\mathcal{A}})$. Estimates of $F_U(\cdot|r)$ can be obtained from the winning bids as in Section 4.2, simply replacing $\mathcal{N}$ with $\mathcal{A}$. Combining such estimators to form

$$\widehat{F}_{U_i}(u) = \bigl[1 - \widehat{F}_{U_i}(r)\bigr]\widehat{F}_{U_i}(u|r) + \widehat{F}_{U_i}(r)$$

leads to a consistent estimator of $F_{U_i}(u)$.

Haile and Tamer (2003) point out that similar extensions apply to the bounds approach to ascending auctions discussed in Section 4.3. Their assumptions (see Section 4.3) imply that all bidders with valuations above the reserve price must participate, as in the standard model. Ignoring the endogenous participation and treating $\mathcal{A}$ as the set of potential bidders then leads to bounds on the CDF $F_U(u|r)$ for $u \geqslant r$. Combining these with an estimate of $F_U(r)$ obtained from the observable participation decision leads to bounds on $F_U(u)$ for $u \geqslant r$.

While we have treated the reserve price above as fixed, it should be clear that this is not necessary. As with other auction-specific covariates, the results above can be interpreted as holding for a given value of the reserve price. However, because economic

theory places considerable structure on the effect of the reserve price on the distribution of participating bidders' valuations, in practice this structure should be utilized in estimation. For example, one would want to use data from all auctions with reserve prices below s to estimate $F_{U_i}(u)$ for $u \geqslant s$. This requires a modified estimation approach that combines data drawn from different truncated distributions. Indeed, if the reserve price varies exogenously (e.g., as it would if it were set optimally by sellers with stochastic private values for the good that are independent of bidders' valuations), this variation can trace out much (or even all) of the distributions $F_{U_i}(\cdot)$. For example, if the support of the reserve price includes values below the lower boundary of the support of bidder valuations, then identification of the full distribution $F_U(\cdot)$ is immediate from the arguments above. The estimation problem in such cases is similar to that for other models with random truncation [e.g., Woodroofe (1985), Wang, Jewell and Tsai (1986)]. While this idea has been mentioned by Guerre, Perrigne and Vuong (2000), nonparametric estimators exploiting the presence of variation in reserve prices have not yet been investigated, either for ascending or first-price auctions.

6.3.1.2. First-price auctions Similar arguments apply to first-price auctions, although here we can consider a richer set of private values models. We will focus on the case in which the econometrician observes all of the bids as well as the realizations of the sets $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{N}}$. In first-price auctions, it is necessary to make an assumption about whether the bidders observe the set $\widetilde{\mathcal{A}}$ before placing their bids. Since participation is determined by the realization of bidders' private information, it will often be most natural to assume that bidders do not know $\widetilde{\mathcal{A}}$ when choosing their bids. We will focus on this case.[64]

Since for any bidder i making a bid in equilibrium

$$G_{M_i|B_i}(m_i|b_i;\mathcal{N}) = \Pr\bigl(\widetilde{\mathcal{A}} = \{i\} \mid i \in \widetilde{\mathcal{A}},\ B_i = b_i, \mathcal{N}\bigr) + \sum_{\mathcal{A}' \subset \mathcal{N}, i \in \mathcal{A}'} \Pr\Bigl(\widetilde{\mathcal{A}} = \mathcal{A}',\ \max_{k \in \mathcal{A}',\, k \neq i} B_k \leqslant m_i \ \Big|\ i \in \widetilde{A},\, B_i = b_i, \mathcal{N}\Bigr)$$

the observables and the first-order condition (2.4) uniquely determine the valuation u_{it} associated with the bid b_{it} of each actual bidder. Letting $F_{\mathbf{U}}(\cdot|\mathcal{A}, r)$ denote the joint distribution of $\{U_i\colon\ U_i \geqslant r,\ i \in \mathcal{A}\}$, this gives the following result.

THEOREM 6.4. *For each $\mathcal{A} \subseteq \mathcal{N}$, the joint distribution $F_{\mathbf{U}}(\cdot|\mathcal{A}, r)$ is identified in a first-price auction from observation of the reserve price r, all bids, and the associated bidder identities. In a symmetric environment, it is sufficient to observe r and all bids.*

Combined with the probabilities $\Pr(\widetilde{\mathcal{A}} = \mathcal{A} \mid \mathcal{N}, r)$ (for which identification is immediate when $\mathcal{A}$, $\mathcal{N}$, and r are all observed), the joint distributions $F_{\mathbf{U}}(\cdot; \mathcal{A}, r)$ will be sufficient for a number of questions of interest, including predicting the effects of an

[64] In some auctions, bidders may be required to register or make a deposit in order to participate. If these actions are observable to bidders, $\mathcal{A}$ will be known at the time they choose their bids.

increase in the reserve price. As discussed above, however, in some cases one will need an estimate of the untruncated distribution of valuations. This does not appear to be possible in the case of correlated private values: there is simply no information available regarding the correlation of valuations below the reserve price. However, maintaining the assumption that $\mathcal{N}$ is observable, one can identify the marginal distributions of bidder valuations evaluated at values above r.[65]

THEOREM 6.5. *In a first-price auction with private values, $F_{U_i}(u_i)$ is identified for all $u_i \geqslant r$ from observation of all bids and the associated bidder identities. In a symmetric environment, it is sufficient to observe all bids.*

PROOF. For each $\mathcal{A}$ and each $i \in \mathcal{A}$, the joint distribution $F_{\mathbf{U}}(\cdot|\mathcal{A}, r)$ completely determines the conditional distribution $F_{U_i}(u_i|r) = \Pr(U_i \leqslant u_i \mid U_i \geqslant r)$. Further,

$$F_{U_i}(u_i|r) = \frac{F_{U_i}(u_i) - F_{U_i}(r)}{1 - F_{U_i}(r)} \tag{6.22}$$

for all $u_i \geqslant r$. $F_{U_i}(r)$ is identified from the observed participation decisions, as in the case of an ascending auction. The result then follows from (6.22). □

Note that in an independent private values auction, this provides identification of $F_{\mathbf{U}}(\mathbf{u})$ for $\mathbf{u}$ such that $u_i \geqslant r$ for all i. As with similar results in preceding sections, estimation is possible building directly on the identification result, substituting sample analogs for the probabilities $F_{U_i}(u_i|r)$ and $F_{U_i}(r)$ in (6.22).

6.3.2. *Costly signal acquisition and the identification of acquisition costs*

Levin and Smith (1994) have considered a model in which players ("firms") first choose whether to become potential bidders ("enter") by investing in signals of their valuations. Firms that invest observe private signals. The assumption of costly signals is natural in many environments, particularly in the procurement contexts that account for a large share of the data studied in the auctions literature. For example, acquiring a signal might require conducting/analyzing a seismic survey or reviewing detailed contract specifications. In this subsection, we discuss identification of both value distributions and the costs of signal acquisition.

Levin and Smith (1994) assume that the bidders observe the set of potential bidders before placing their bids; in Section 6.3.3 we discuss the alternative assumption that investments in signal acquisition are private information so that bidders place their bids without knowing which firms are potential bidders. Levin and Smith focus on symmetric

[65] Analogs of Theorems 6.4 and 6.5 were demonstrated for the case of symmetric independent private values by Guerre, Perrigne and Vuong (2000). Haile, Hong and Shum (2003) extended these results to symmetric affiliated private values and common values models.

equilibria of models with symmetric bidders. In equilibrium, firms acquiring a signal must expect to recover the cost of doing so on average. So when there are sufficiently many firms in the market, some must choose not to enter. In the unique symmetric equilibrium, entry is determined by mixed strategies, leading to exogenous variation in the set of potential bidders.[66] One caveat is that, as in virtually all entry games, when asymmetric equilibria are allowed, there will be multiple equilibria [see, e.g., Berry and Tamer (2005)].

To extend the econometric model to this setting, observe that the distribution of the set of potential bidders is determined by the mixing probabilities. Since firms make independent decisions about signal acquisition, the event $|\widetilde{\mathcal{N}}| = 1$ occurs with positive probability. This case was ruled out above because typically this is not an interesting case: if a bidder knows that $|\mathcal{N}| = 1$, she will simply bid the reserve price. However, for the purposes of this section and the next, we will allow $|\mathcal{N}| = 1$. If we assume that the reserve price r is less than $\underline{u}_i$ for all firms i, the reserve price plays a role only when $|\mathcal{N}| = 1$, in which case the lone potential bidder bids the reserve. Hence when $r < \underline{u}_i$ for all i, the number of potential bidders is equal to the number of actual bidders, the model generates exogenous variation in the number of bidders, and the methods described above can be used to estimate primitive value distributions. When $r > \underline{u}_i$ there will also be also variation in the number of actual bidders for a given set of potential bidders, as in Section 6.3.1. There we assumed that the set of potential bidders was observable to the econometrician for some results. That may be unlikely in the presence of both a reserve price and costly signals, since the set of potential bidders varies across auctions. Li (2003) considers parametric estimation of a model based on Levin and Smith's model with $r > \underline{u}_i$.

In their study of US Forest Service timber auctions, Athey, Levin and Seira (2004) consider a variation of this model, allowing asymmetric bidders. They assume firms fall into two classes, "weak" and "strong" (generalizations to more than two types are also possible). Strong firms that choose to invest draw valuations from a distribution that stochastically dominates that of the "weak" firms. They restrict attention to type-symmetric equilibria, in which all members of a given class use the same strategies. Because firms are asymmetric, however, there may be multiple type-symmetric equilibria. Athey, Levin and Seira (2004) derive a restriction on primitives that guarantees a unique type-symmetric equilibrium, and this restriction can be verified empirically.

In any signal acquisition model that generates exogenous variation in $\widetilde{\mathcal{N}}$, if $\widetilde{\mathcal{N}}$ and all bids are observed (or in an IPV model if $\widetilde{\mathcal{N}}$ and the winning bid are observed), our prior results imply that (assuming $r < \underline{u}_i$) a bidder's *ex ante* gross expected profit $\Pi_i(\mathcal{N})$ from entering the auction is identified. In particular,

$$\Pi_i(\mathcal{N}) = E_{U_i}\left[\left(U_i - \beta_i(U_i;\mathcal{N})\right)G_{M_i|B_i}\left(\beta_i(U_i;\mathcal{N}) \mid \beta_i(U_i;\mathcal{N});\mathcal{N}\right)\right]$$

[66] Hendricks, Pinkse and Porter (2003) have considered a variation on this model in a common values setting in which bidders choose whether to invest in a signal based on noisier (in a precise sense) preliminary estimates of their valuations. As they point out, their model can be interpreted as providing a purification of Levin and Smith's (1994) mixed strategy equilibrium.

with the right-hand side determined by the observed bid distribution and the first-order conditions for equilibrium bidding. Identification of $\Pi_i(\mathcal{N})$ requires no assumptions about the nature of the signal acquisition equilibrium (or equilibrium selection) beyond what is required to guarantee that variation in $\widetilde{\mathcal{N}}$ is exogenous. Estimates of $\Pi_i(\mathcal{N})$ can then be used to calculate all equilibria of an entry game for given entry costs. Thus, in an application, the existence of multiple equilibria in the entry game can be assessed empirically.

Athey, Levin and Seira (2004) show that in the unique type-symmetric equilibrium in their application, strong firms enter with probability one and weak firms are indifferent about entry. They further observe that for any firms that are indifferent about acquiring a signal (the weak firms in their application), the expected profit from entry must be zero. Thus entry costs are identified using $\Pi_i(\mathcal{N})$ and the distribution of $\widetilde{\mathcal{N}}$, which is directly observable. In particular, for any firm i that is indifferent about acquiring a signal, signal acquisition costs must be equal to

$$\overline{\Pi}_i = \sum_{\mathcal{N}:\, i \in \mathcal{N}} \Pr(\widetilde{\mathcal{N}} = \mathcal{N} \mid i \in \widetilde{\mathcal{N}}) \Pi_i(\mathcal{N}).$$

Thus, in contrast to much of the empirical industrial organization literature on entry (where entry corresponds to signal acquisition in this model), which draws inferences solely from entry decisions,[67] the level of entry costs can be inferred. Hence it is possible to conduct counterfactual simulations about changes in these costs on the competitiveness of markets and bidder rents.

6.3.3. Bidder uncertainty about the competition

Throughout the preceding sections we maintained the assumption that bidders make their bids knowing the set of competitors they face. In the standard model of the ascending auction with private values, this is without loss of generality since the dominant strategy is not affected by the set of opponents. Furthermore, the assumption may be uncontroversial in an ascending auction; certainly if one believes bidders observe their opponents' exit prices (as in the standard model) it is natural to presume that bidders are aware of all competitors. In a sealed-bid auction, however, bidders need not gather to participate, making it less certain that bidders will know what competition they face. And in a first-price auction, a bidder's information about the competition is critical to the characterization of equilibrium bidding. In some procurement settings, firms may in fact know which of their competitors have the capability to compete for a given contract or even which firms have been invited to bid, but in other contexts this may not be public information.

Even if the set of firms who could in principle compete in an auction is common knowledge, in models where firms incur a cost to acquire a signal (see, e.g., Section 6.3.2) bidders may not know which other firms have actually invested in a signal

[67] See, e.g., Berry and Reiss (in press).

for a particular auction. There the investment choice is determined by randomization (in the case of a mixed strategy equilibrium) or as a function of private information (in a pure strategy equilibrium).

It is straightforward to modify theoretical models of costly signal acquisition to accommodate the case where bidders do not observe who has acquired a signal before bidding. McAfee, Quan and Vincent (2002) and Hendricks, Pinkse and Porter (2003) consider models with this feature for the case of first-price auctions. McAfee, Quan and Vincent (2002) show that a slightly stronger condition than affiliation of signals is required to ensure existence of a pure strategy Nash equilibrium in increasing strategies.[68] Li and Zheng (2005) also study such a model, highlighting an interesting testable theoretical possibility: bids may decrease when the number of firms increases, because each firm will enter with lower probability, and the resulting change in the distribution of potential bidders has ambiguous consequences for bidding strategies.

6.3.3.1. Unknown potential competition Now consider a first-price sealed-bid auction where $\widetilde{\mathcal{N}}$ is unobserved to both bidders and the econometrician. $\Pr(\widetilde{\mathcal{N}} = \mathcal{N})$ is identified as long as the set of bidders is observable at each auction. The distribution of the highest bid among i's opponents is calculated taking the expectation over the set of potential bidders:

$$\begin{aligned} &G_{M_i|B_i}(m_i|b_i) \\ &\quad = \Pr\bigl(\widetilde{\mathcal{N}} = \{i\} \mid i \in \widetilde{\mathcal{N}}\bigr) \\ &\qquad + \sum_{\mathcal{N}:\, i\in\mathcal{N},\, |\mathcal{N}|>1} \Pr\Bigl(\max_{j\in\mathcal{N},\, j\neq i} B_j \leqslant m_i \;\Big|\; i \in \widetilde{\mathcal{N}}, B_i = b_i\Bigr) \Pr(\widetilde{\mathcal{N}} = \mathcal{N} \mid i \in \widetilde{\mathcal{N}}). \end{aligned} \tag{6.23}$$

Bidder i's first-order condition is then given by

$$u_i = b_i + \frac{G_{M_i|B_i}(b_i|b_i)}{g_{M_i|B_i}(b_i|b_i)}. \tag{6.24}$$

This takes the usual form; however, here $G_{M_i|B_i}(\cdot)$ does not depend on $\mathcal{N}$. Equation (6.24) and observation of all bids then identifies the distribution of U_i, and straightforward extensions of the estimation techniques described above can be applied.

So far we have considered two assumptions that might be made when interpreting data from first-price auctions: (i) $\widetilde{\mathcal{N}}$ is observed by all bidders prior to bidding, or (ii) $\widetilde{\mathcal{N}}$

[68] In particular, they assume that there exists a nondecreasing function $h(\cdot)$ such that for each i, $U_i = h(X_i, V_0)$, where $(X_1, \ldots, X_n)$ are i.i.d. conditional on V_0. Private values, where $U_i = X_i$, is a special case. Each bidder bears a cost (constant across bidders) to learn the value of X_i. Bidders first invest in their signals and then place bids, but investment decisions are not publicly observable. They derive an equilibrium in which firms randomize in the signal acquisition decision. Then, for bidders who acquire a signal, bidding is in strictly monotone pure strategies. They show that a sufficient (but not necessary) condition for existence of a such an equilibrium is that $1 - \rho(1 - F_{X_i|V_0}(x_i|v_0))$ is log-supermodular in (x_i, v_0), where ρ is the entry probability in the mixed strategy equilibrium of the entry game.

is unobserved prior to bidding. In many settings, institutional detail may be available to guide the choice between these assumptions. When there is variation in $\mathcal{N}$, the data can also help guide this choice. If $\widetilde{\mathcal{N}}$ is observed by all bidders prior to bidding, then when $|\mathcal{N}| = 1$ the bidder must bid the reserve price. Thus, the hypothesis that $\widetilde{\mathcal{N}}$ is observable to bidders could be rejected if $\Pr(B^{(1:1)} = r \mid |\widetilde{\mathcal{N}}| = 1) < 1$. In addition, building on the discussion in Section 5, we note that both assumptions can have additional testable implications. If variation in $\widetilde{\mathcal{N}}$ is exogenous (as in the models of costly signal acquisition described above), it will be possible to estimate $F_{\mathbf{U}}(\cdot)$ separately for each $\mathcal{N}$ and compare the resulting estimates. If $\widetilde{\mathcal{N}}$ is observed by bidders before bidding, these estimates should be equal to each other (up to sampling error). On the other hand, if bidders have no information regarding the realization of $\widetilde{\mathcal{N}}$ when choosing their bids, then the distribution of B_i itself should not vary with $\mathcal{N}$ (recall (6.24)).

6.3.3.2. Noisy knowledge of the competition Once we allow the possibility that bidders do not observe $\widetilde{\mathcal{N}}$ prior to bidding, it is natural to consider more carefully what bidders do know. In particular, it may be more reasonable to imagine that bidders have noisy signals of $\widetilde{\mathcal{N}}$ when choosing their bids. When the econometrician can condition on the same information available to bidders (excluding their signals of course), extending the methods is straightforward. Suppose, for example, that bidders form their beliefs about the set of competitors based on a public signal η that is also observable to the econometrician. The signal, η, might contain information about how costly it will be to evaluate the object and acquire a signal, or information about the expected value of the object. In a model of costly signal acquisition, such factors will affect the entry probability of each bidder.

We can extend the methods above by treating η as an auction-specific covariate to be conditioned on in bidders' first-order conditions. Note that the signal η need not be a scalar and can include any information that may affect the set of potential bidders, including, e.g., characteristics of the good for sale or market conditions. Hendricks, Pinkse and Porter (2003) consider a simple example of this approach. They construct a binary signal $\eta_t = 1\{\Upsilon_t \geqslant \Upsilon^*\}$ of the number of potential bidders for tract t, where Υ_t is the number of firms ever to bid on an oil tract in a geographic neighborhood of the tract offered in auction t, and Υ^* is a specified threshold value.

In contrast, if bidders have signals (public or private) about factors that affect the number of competitors, but these signals are not observable to the econometrician, the problem of unobserved heterogeneity discussed in Section 6.1.2 arises. For example, in a model where acquiring a signal is costly, firms might observe an auction characteristic v_0 *before* making entry decisions. Another possibility is that firms observe auction characteristics that affect the cost of acquiring information about a particular object. Li and Zheng (2005) develop a model of a first-price auction with these features. They specify a semi-parametric model, leaving the distribution of unobserved heterogeneity unrestricted while assuming a functional form for the marginal distributions of valuations conditional on the heterogeneity. They estimate the model using Bayesian methods.

6.3.4. Internet auctions and unobserved participation

Internet auctions have recently attracted considerable attention from economists. In addition to providing a great deal of new data, Internet auctions introduce a number of new and interesting questions, including the role of seller reputations [see, e.g., the papers surveyed by Bajari and Hortaçsu (2003b, 2004)] and competition between sellers [e.g., Peters and Severinov (2006)].

Internet auctions are most often conducted in one of several variations on the standard ascending auction mechanism [Lucking-Reiley (2000)]. However, a challenge to structural analysis of bid data from Internet auctions is the fact that the number of bidders cannot be observed. Recall from Section 4 that a key assumption for the identification arguments in even the simplest ascending auction environments was observation of the number of bidders – either the number of potential bidders or the number who have valuations above the reserve price. An Internet auction typically takes place over several days (usually a week or more on eBay, for example), with bidders becoming aware of the auction at different times as they log onto the auction site while the auction is underway. A bidder who logs on to discover that the price has already risen past his valuation will not bid. Hence the number of submitted bids will not generally equal the number of bidders willing to pay the reserve price (if any).[69] The usual assumption that the transaction price is equal to the second-highest valuation is of little use if it is not known whether it is the second highest of two valuations or of ten, for example.

This problem has accounted for a substantial impediment to progress in addressing questions about the underlying demand structures at Internet auctions.[70] This includes even seemingly simple questions like how seller reputations affect bidders' willingness to pay, since this requires inference on the underlying distribution of bidder valuations.

Song (2003) has proposed a model capturing key departures of Internet auctions from the standard ascending auction model. Using this model, she derives conditions under which the identification of the distribution $F_U(\cdot)$ can be obtained in the symmetric independent private values paradigm without observing the number of bidders, or even assuming that this number is constant.

In her model, an auction takes place over an interval of time $[0, \tau]$. The distribution of N can vary across auctions, and need not be known to bidders. In a given auction, each potential bidder i draws a vector of "bidding opportunities" $(t_i^1, \ldots, t_i^{\tau_i})$, with each $t_i^k \in [0, \tau]$. Taking $t_i^1 < \cdots < t_i^{\tau_i}$ without loss of generality, t_i^1 represents the time of i's "arrival" at the auction, and $t_i^{\tau_i}$ represents i's final bidding opportunity. No restriction is

[69] This problem can also arise in other applications, particularly in other ascending auctions with similar deviations from the button auction model, or in Dutch auctions, where only the winner makes a bid. Song (2004) explores identification and estimation in these and other auction models when the number of bidders is not observable to the econometrician.

[70] Bajari and Hortaçsu (2003a) avoid this problem with a common values model that admits an equilibrium in which all bidders willing to pay the reserve price will bid simultaneously at the end of the auction, as if in a second-price sealed bid auction. See Ockenfels and Roth (2006) for an alternative model of Internet auctions.

placed on the joint distribution of $(N, \{\tau_i\}, \{t_i^k\})$ except that (a) these are independent of bidders' valuations, and (b) each $t_i^{\tau_i}$ is continuously distributed on some interval $(t_i^0, \tau]$. In this model, bidders may "arrive" early or late, bid frequently or infrequently, and have different notions of what bidding at the "last minute" means.

At each bidding opportunity, a bidder may specify a "cutoff price" of any value above the current standing bid. Whenever a new cutoff price is submitted, the auctioneer raises the standing bid (denoted s_t) to the second-highest cutoff price, and the bidder with the highest cutoff price is named the standing high bidder. This matches the actual procedure on eBay, the most popular Internet auction site, for example. At time τ, the standing high bidder wins the object at the standing bid (for simplicity we assume no reserve price and no minimum bid increment). Typically, the econometrician can observe the history of submitted cutoff prices (except the winner's), as well as the identity of the bidder who placed each bid. This information is publicly available for eBay auctions, for example.

There are many equilibria of this game. For example, all bidders can submit cutoff prices equal to their valuations at their first bidding opportunities; bidders may start with low cutoff prices and gradually raise them as the auction proceeds; or some/all bidders may wait until their final bidding opportunities to submit a cutoff price. In some of these equilibria (like the last example), some potential bidders will not bid, since at their planned bidding time the standing bid will already exceed their valuations. However, Song (2003) shows that in all equilibria the highest cutoff price submitted by bidder i will be no larger than his valuation u_i, and it will equal his valuation if the standing bid at time $t_i^{\tau_i}$ was below u_i.[71]

Since the price can never rise above $u^{(n-1:n)}$, this ensures that the allocation is efficient and that the transaction price is $u^{(n-1:n)}$. Further, in some cases, the third-highest cutoff price submitted will be equal to $u^{(n-2:n)}$. To see this, let b_i denote the highest cutoff price submitted by bidder i (i.e., his "bid") and let $b^{(m-2:m)}$ denote the third-highest such bid (or $-\infty$ if there is no such bid). Here m represents the number of *observed* bidders – those submitting cutoff prices at some point in the auction. Now suppose that at time $\tilde{t}$ the standing bid $s_{\tilde{t}}$ is no higher than $b^{(m-2:m)}$. In practice, whether this is true can be directly determined from the available bidding histories. In particular, recalling that two bids above b are required for the standing bid to exceed b, this occurs whenever at least one of the two bidders making the highest bids (as of the end of the auction) makes no bid above $b^{(m-2:m)}$ prior to time $\tilde{t}$. In that case we have

$$s_{\tilde{t}} \leqslant b^{(m-2:m)} \leqslant u^{(m-2:m)} \leqslant u^{(n-2:n)}$$

implying that if the bidder with valuation $u^{(n-2:n)}$ has his final bidding opportunity at time $\tilde{t}$ or earlier, he will submit a cutoff price equal to his valuation. In that case,

[71] The first property is easily understood. The second follows from arguments similar to those used in analyzing a second-price auction.

$b^{(m-2:m)} = u^{(n-2:n)}$. While the final bidding time of this bidder is not known, by looking at auctions in which $s_{\tilde{t}} \leqslant b^{(m-2:m)}$ for $\tilde{t}$ sufficiently close to τ, the probability that $b^{(m-2:m)} = u^{(n-2:n)}$ can be made arbitrarily close to one.[72]

This is useful because we typically cannot observe the cutoff price submitted by the auction winner.[73] However, by examining only the set of auctions in which

$$s_{\tilde{t}} \leqslant b^{(m-2:m)}, \quad \tilde{t} \in (\tau - \delta, \tau), \tag{6.25}$$

for small $\delta > 0$, we can treat the order statistics $(U^{(n-1:n)}, U^{(n-2:n)})$ as "observed." The following result, proved in Song (2003), implies that observation of $(U^{(n-1:n)}, U^{(n-2:n)})$ is sufficient to identify $F_U(\cdot)$, even though the realization of N at each auction is unknown.

LEMMA 6.2. *Let $(Y^{(N:N)}, Y^{(N-1:N)}, Y^{(N-2:N)})$ denote random variables equal to the three highest of $N \geqslant 3$ independent draws from a univariate distribution $F_Y(\cdot)$, where N is stochastic and unobserved. $F_Y(\cdot)$ is uniquely determined by the joint distribution of $(Y^{(N-1:N)}, Y^{(N-2:N)})$.*

PROOF. Given $Y^{(N-2:N)} = y'$, the pair $(Y^{(N-1:N)}, Y^{(N:N)})$ can be reinterpreted as the two order statistics for an i.i.d. sample of size two from the distribution

$$F_Y(\cdot|y') = \frac{F_Y(\cdot) - F_Y(y')}{1 - F_Y(y')}.$$

Although $Y^{(N:N)}$ is unobserved, Equation (4.1) implies that the observation of $Y^{(N-1:N)}$ alone is sufficient to identify the parent distribution $F_Y(\cdot|y')$ for this sample. Identification of $F_Y(\cdot)$ then follows from the fact that

$$\lim_{y' \downarrow \inf \operatorname{supp} F_Y^{(N-2:N)}(\cdot)} F_Y(\cdot|y') = F_Y(\cdot).$$

□

Key to the applicability of this result is an assumption that auctions in which at least one of the two high bidders make late bids (i.e., where (6.25) holds) are representative of all auctions. If $\tau_i = 1\ \forall i$, this follows from the assumption that $\{\tau_i\}$ and $\{U_i\}$ are independent. In general, when $\tau_i > 1$, this requires the additional assumption that the equilibrium selection does not depend on $(u_1, \ldots, u_n)$.[74]

[72] In finite sample, of course, there will be a tradeoff between the bias of including auctions with $\tilde{t}$ far from τ and the reduction in the variance from doing so. Song (2003) suggests a data-driven approach for choosing the sample.

[73] If we could – for example, if such data were provided directly by eBay – a variation on the result below would still be applicable to address the problem that the number of potential bidders is unobserved.

[74] More precisely, the distribution of $U^{(n-1:n)}|U^{(n-2:n)}$ conditional on at least three bidders' being observed and (6.25) holding must be the same as the unconditional distribution of $U^{(n-1:n)}|U^{(n-2:n)}$.

Song (2003) proposes a semi-nonparametric estimator [Gallant and Nychka (1987)] applicable to the subset of auctions in which bids are observed from at least three distinct bidders. The likelihood function is constructed from the conditional density of $U^{(n-1:n)}$ given $U^{(n-2:n)}$, i.e.,

$$\frac{\partial}{\partial y}\Pr\left(U^{(n-1:n)} \leqslant y \mid U^{(n-2:n)} = x\right) = \frac{2(1 - F_U(y))\, f_U(y)}{(1 - F_U(x))^2}$$

in which n does not appear. Monte Carlo experiments suggest that the approach can perform well in sample sizes easily attainable from Internet auctions.

6.4. Risk aversion

Most of the empirical literature on auctions assumes risk neutrality of bidders. Risk neutrality is a natural assumption when the value of the object being sold is small relative to each bidder's wealth. Furthermore, in many applications bidders are firms, which economists usually assume to be profit maximizers. However, many auctions involve highly valuable goods (or contracts). And even when bidders represent firms, they may themselves be risk averse.[75] Risk aversion can have important implications for a wide range of policy questions, including the optimal reserve price and a seller's preference between the standard auction formats.[76]

Risk aversion also creates significant challenges for identification. In an ascending auction with private values, for example, risk aversion has no effect on equilibrium bidding in the standard model: bidding one's valuation is still a dominant strategy. While this implies that identification of $F_{\mathbf{U}}(\cdot)$ holds with risk aversion whenever it holds with risk neutrality, it also implies that there is no way to distinguish risk neutrality from risk aversion, i.e., no way to identify bidders' preferences. While the distribution $F_{\mathbf{U}}(\cdot)$ will be sufficient for some questions (for example, the effect of changing the reserve price) it will be inadequate for many others.

In a first-price auction, the implications of risk aversion for equilibrium bidding are nontrivial, since bidding involves a gamble. Bidding less aggressively leads to a lower chance of winning but higher profits conditional on winning. A more risk averse bidder will be less willing to accept a reduced probability of winning in order to obtain a higher profit when she wins. This suggests that there is at least hope for identification of preferences using data from first-price auctions. However, identifying risk preferences generally requires observation of choices from different menus of lotteries. Variations

[75] The incentives provided by the firms they work for may or may not "undo" such risk aversion. Athey and Levin (2001), Campo et al. (2002), and Perrigne (2003), for example, find evidence consistent with risk averse bidding behavior by firms at timber auctions.

[76] See, e.g., McAfee and McMillan (1987) and references therein. The theory of first-price auctions with risk averse bidders was initially developed by Maskin and Riley (1984). Campo et al. (2002) extend the analysis to the case in which there is no binding reserve price and establish additional smoothness properties used for identification and estimation.

in bidders' valuations do change the sets of lotteries available to them, but not in ways that are observable to the econometrician, since valuations are private information. This suggests that some observable exogenous variation will be needed to separately identify preferences and the distribution of valuations.[77] Below we will first explore possible approaches to identification in symmetric models, proceeding to consider models with asymmetric preferences in Section 6.4.2.

6.4.1. Symmetric preferences

We begin with a more formal illustration of the fundamental challenge for identification in models of first-price auctions with risk aversion. For simplicity, consider the case of symmetric independent private values, and assume that all bids are observable. Assume further that all bidders share the same continuously differentiable utility function $\omega(\cdot)$. Taking equilibrium behavior of her opponents as given, bidder i solves the problem

$$\max_{\tilde{b}_i} \omega(u_i - \tilde{b}_i) \Pr\Big(\max_{j \in \mathcal{N}_{-i}} B_j \leqslant \tilde{b}_i \Big).$$

If we define

$$\lambda(s) \equiv \omega(s)/\omega'(s), \tag{6.26}$$

then first-order condition

$$\omega'(u_i - b_i) G_B(b_i) = (n-1)\omega(u_i - b_i) g_B(b_i)$$

can be rewritten usefully as

$$u_i = b_i + \lambda^{-1}\left(\frac{1}{n-1} \frac{G_B(b_i)}{g_B(b_i)} \right). \tag{6.27}$$

Now define the function

$$\xi(b_i, n, \lambda) = b_i + \lambda^{-1}\left(\frac{1}{n-1} \frac{G_B(b_i)}{g_B(b_i)} \right)$$

and let $\lambda_{\mathrm{I}}(\cdot)$ denote the identity function – i.e., the function $\lambda(\cdot)$ implied in the case of risk neutrality. Campo et al. (2002) show that as long as bids are independent (and additional regularity conditions are satisfied), an observed marginal bid distribution $G_B(\cdot)$ can be rationalized by equilibrium behavior if and only if there exists a utility function $\omega(\cdot)$ such that, for the associated $\lambda(\cdot)$, $\xi(\cdot; n, \lambda)$ is increasing (see Section 5.1). Hence, if bids are independent and $\xi(\cdot, n, \lambda_{\mathrm{I}})$ is increasing, it will be possible to find a distribution $F_U(\cdot)$ that rationalizes the observed bids within the symmetric risk neutral IPV

[77] However, if the distribution $F_{\mathbf{U}}(\cdot)$ is known or identified from other data – for example, in a laboratory setting or when one observes the same bidders participating in both first-price and ascending auctions – bid data from first-price auctions might then be used to estimate the utility function.

model. If $\xi(\cdot, n, \lambda_I)$ is decreasing at some point, the observed bids could not have been generated by equilibrium bidding by risk-neutral bidders, although there may exist another utility function $\omega(\cdot)$ with associated $\lambda(\cdot)$ such that $\xi(\cdot; n, \lambda)$ is increasing. Thus, allowing for risk aversion expands the set of observable bid distributions that can be rationalized by equilibrium bidding [Campo et al. (2002)].

Unfortunately, $\xi(\cdot, n, \lambda)$ need not violate the monotonicity restriction when the model is misspecified – in particular when the given function $\lambda(\cdot)$ does not correspond to that for the true preferences. When $\xi(\cdot, n, \lambda_I)$ is increasing, for example, the observed bids can be rationalized with risk neutrality, but they can also be rationalized with many different specifications of risk aversion. To suggest why, observe that if as long as a bidder is sufficiently risk averse, $\lambda^{-1}(\frac{1}{n-1}\frac{G_B(b)}{g_B(b)})$ does not vary much with b, ensuring that $\xi(b; n, \lambda)$ strictly increases in b. Consider the following example of a CRRA utility function (with zero initial wealth): $\omega(u) = u^{1-c}$, with $0 \leqslant c < 1$. Then $\lambda(s) = s/(1-c)$, and $\lambda^{-1}(z) = z(1-c)$. As c approaches 1, $\xi(\cdot; n, \lambda)$ approaches the identity function. Intuitively, sufficiently risk averse bidders are not willing to risk losing the object by shading their bids, so they do not respond to the shape of the opposing bid distribution. Thus, even if $\frac{G_B(b)}{g_B(b)}$ is sharply decreasing in some places, there will be a critical level of risk aversion above which $\xi(\cdot; n, \lambda)$ is everywhere increasing.

Similarly, it is generally impossible to identify the degree of risk aversion from bid data in a fixed environment. Perhaps surprisingly,[78] however, this is true even with a strong functional form assumption on bidders' preferences. Again consider the CRRA example, and suppose that the data can be rationalized by a distribution $F_U(\cdot)$ and coefficient of relative risk aversion c. Then, for any $\tilde{c} \in (c, 1)$, if we let $\tilde{\omega}(s) = s^{1-\tilde{c}}$, we can find another distribution $F_{\widetilde{U}}(\cdot)$ that implies the same distribution of bids, but where $F_{\widetilde{U}}(\cdot)$ stochastically dominates $F_U(\cdot)$. In particular, to satisfy (6.27), we define $\widetilde{U}_i$ to be equal in distribution to

$$\begin{aligned}
\xi(B_i; n, \tilde{\lambda}) &= B_i + \frac{1-\tilde{c}}{n-1}\frac{G_B(B_i)}{g_B(B_i)} \\
&= \frac{\tilde{c}-c}{1-c}B_i + \frac{1-\tilde{c}}{1-c}\left(B_i + \frac{1-c}{n-1}\frac{G_B(B_i)}{g_B(B_i)}\right) \\
&= \frac{\tilde{c}-c}{1-c}B_i + \frac{1-\tilde{c}}{1-c}\xi(B_i; n, \lambda) \\
&= \frac{\tilde{c}-c}{1-c}B_i + \frac{1-\tilde{c}}{1-c}U_i.
\end{aligned}$$

It follows that whenever $\xi(b_i; n, \lambda)$ is increasing in b_i, so is $\xi(b_i; n, \tilde{\lambda})$. Hence, the data can be rationalized with risk aversion $\tilde{c}$. Campo et al. (2002) show that this argument holds for other parameterized families, as well as general utility functions.

[78] Recall that with risk neutrality the symmetric IPV model is overidentified when one observes all bids from each auction.

These results are quite negative. Only for bid distributions such that $\frac{G_B(b)}{g_B(b)}$ decreases sufficiently sharply in b in places can risk aversion be distinguished from risk neutrality; it is impossible to distinguish among different parameterized functional forms for risk aversion; and there exists a large range of risk aversion parameters that can rationalize the observed bid data, even when attention is restricted to a particular functional form.

Following the intuition at the beginning of this section, however, this nonidentification might be overcome with observable exogenous variation in the sets of gambles available to bidders. One possibility is a covariate that shifts bidders' initial wealth or, equivalently, bidders' valuations for the good. Suppose, for example, that each bidder i's utility from winning the auction is

$$\omega\big(h(w_i) + u_i - b_i\big) \tag{6.28}$$

for some increasing function $h(\cdot)$, where the covariate w_i is independent of u_i and is observable to all bidders prior to the auction as well as to the econometrician. Let $\mathbf{w} = (w_1, \ldots, w_n)$. The model then becomes asymmetric, even though bidders' preferences are given by the same function. Let $G_{M_i}(b_i|\mathbf{w}, \mathcal{N})$ be the distribution of the maximum bid of bidder i's opponents, conditional on $\mathbf{w}$ and $\mathcal{N}$. Let $b_{\alpha,\mathbf{w},\mathcal{N}}$ denote the αth quantile of the distribution $G_{M_i}(b_i|\mathbf{w}, \mathcal{N})$, while u_α is the αth quantile of $F_U(\cdot)$. Then equilibrium requires that

$$\begin{aligned} u_\alpha &= b_{\alpha,\mathbf{w},\mathcal{N}} - h(w_i) + \lambda^{-1}\left(\frac{G_{M_i}(b_{\alpha,\mathbf{w},\mathcal{N}}|x, \mathcal{N})}{g_{M_i}(b_{\alpha,\mathbf{w},\mathcal{N}}|x, \mathcal{N})}\right) \\ &\quad \forall \mathbf{w} \in \operatorname{supp} \mathbf{W},\ \forall \alpha \in [0, 1]. \end{aligned} \tag{6.29}$$

The data can be rationalized by the model only if we can find a $\lambda(\cdot)$ such that (6.29) holds and such that

$$\xi_i(b_i; \mathcal{N}, \lambda, \mathbf{w}) \equiv b_i - h(w_i) + \lambda^{-1}\left(\frac{G_{M_i}(b_i|\mathbf{w}, \mathcal{N})}{g_{M_i}(b_i|\mathbf{w}, \mathcal{N})}\right)$$

is increasing in b_i.

This may not be possible, especially within a restricted class of utility functions. To see this, again consider the one-parameter CRRA example and suppose that the vector $\mathbf{W}$ takes on only two values, $\mathbf{w}'$ and $\mathbf{w}''$. Then equilibrium requires that for all $\alpha \in [0, 1]$,

$$\begin{aligned} u_\alpha &= b_{\alpha,\mathbf{w}'',\mathcal{N}} - h\big(w_i''\big) + (1 - c)\frac{G_{M_i}(b_{\alpha,\mathbf{w}'',\mathcal{N}}|\mathbf{w}'', \mathcal{N})}{g_{M_i}(b_{\alpha,\mathbf{w}'',\mathcal{N}}|\mathbf{w}'', \mathcal{N})} \\ &= b_{\alpha,\mathbf{w}',\mathcal{N}} - h\big(w_i'\big) + (1 - c)\frac{G_{M_i}(b_{\alpha,\mathbf{w}',\mathcal{N}}|\mathbf{w}', \mathcal{N})}{g_{M_i}(b_{\alpha,\mathbf{w}',\mathcal{N}}|\mathbf{w}', \mathcal{N})}. \end{aligned}$$

Thus

$$c = 1 - \frac{b_{\alpha,\mathbf{w}'',\mathcal{N}} - b_{\alpha,\mathbf{w}',\mathcal{N}} - (h(w_i'') - h(w_i'))}{\frac{G_{M_i}(b_{\alpha,\mathbf{w}',\mathcal{N}}|\mathbf{w}',\mathcal{N})}{g_{M_i}(b_{\alpha,\mathbf{w}',\mathcal{N}}|\mathbf{w}',\mathcal{N})} - \frac{G_{M_i}(b_{\alpha,\mathbf{w}'',\mathcal{N}}|\mathbf{w}'',\mathcal{N})}{g_{M_i}(b_{\alpha,\mathbf{w}'',\mathcal{N}}|\mathbf{w}'',\mathcal{N})}}. \tag{6.30}$$

For a given quantile α, rationalizing the data with the CRRA model requires that there exist a function $h(\cdot)$ such that this c lies in the interval $[0, 1)$. If no such $h(\cdot)$ exists, then we can immediately reject the CRRA model. Of course, (6.30) must hold for all quantiles α. Unless the ratio on the right side of (6.30) is invariant to the quantile α, the model will be rejected. Thus, a more flexible specification of risk preferences will typically be required to rationalize the observed bidding data when there are bidder-specific covariates shifting wealth or valuations.

Of course, there is more than one way to relax the structure imposed by (6.28) and CRRA. Campo et al. (2002) maintain the CRRA specification above but assume a functional form for the effect of covariates on valuations only at a single quantile of the distribution of valuations. By leaving the effects at other quantiles unspecified, the problem that the data may reject the model is avoided. We refer readers to their paper for details, as well as an estimation approach.[79]

Another possible approach to identification is to exploit exogenous variation in the number of bidders [e.g., Bajari and Hortaçsu (2005)]. Such exogenous variation changes the equilibrium probability that each given bid wins and, therefore, changes the lotteries available to bidders. Note that unlike the effect of a covariate on the utility gain from winning, this variation in the probability of winning can be determined directly from the equilibrium bid distribution for each $\mathcal{N}$. Using the CRRA model as an example, suppose that there are two groups of bidders, $\mathcal{N}$ and $\mathcal{N}'$, with $|\mathcal{N}| = n$ and $|\mathcal{N}'| = n+1$. Letting $b_{\alpha,\mathcal{N}}$ be the αth quantile of $G_{B_i}(b_i|\mathcal{N})$, equilibrium requires

$$u_\alpha = b_{\alpha,\mathcal{N}} + \frac{1-c}{n-1}\frac{G_{B_i}(b_{\alpha,\mathcal{N}}|\mathcal{N})}{g_{B_i}(b_{\alpha,\mathcal{N}}|\mathcal{N})} = b_{\alpha,\mathcal{N}'} + \frac{1-c}{n}\frac{G_{B_i}(b_{\alpha,\mathcal{N}'}|\mathcal{N}')}{g_{B_i}(b_{\alpha,\mathcal{N}'}|\mathcal{N}')}$$

so that

$$c = 1 - \frac{b_{\alpha,\mathcal{N}'} - b_{\alpha,\mathcal{N}}}{\frac{1}{n-1}\frac{G_{B_i}(b_{\alpha,\mathcal{N}}|\mathcal{N})}{g_{B_i}(b_{\alpha,\mathcal{N}}|\mathcal{N})} - \frac{1}{n}\frac{G_{B_i}(b_{\alpha,\mathcal{N}'}|\mathcal{N}')}{g_{B_i}(b_{\alpha,\mathcal{N}'}|\mathcal{N}')}}.$$

Again, for a given α, this c may not lie in $[0, 1)$ and, further, the right-hand side may not be constant in α as required. Thus, exogenous variation in participation may allow us to reject the CRRA model (or other parameterized utility functions). This suggests some hope of identifying preferences. In the completely general case, one would need to find a utility function such that

$$b_{\alpha,\mathcal{N}} + \lambda^{-1}\left(\frac{1}{n-1}\frac{G_{B_i}(b_{\alpha,\mathcal{N}}|\mathcal{N})}{g_{B_i}(b_{\alpha,\mathcal{N}}|\mathcal{N})}\right)$$

is invariant to $\mathcal{N}$ for each α. Depending on how much variation in $b_{\alpha,\mathcal{N}}$ and $\frac{G_{B_i}(b_{\alpha,\mathcal{N}}|\mathcal{N})}{g_{B_i}(b_{\alpha,\mathcal{N}}|\mathcal{N})}$ is induced by variation in $\mathcal{N}$ and α, it may be possible to identify the entire utility function.

[79] Bajari and Hortaçsu (2005) propose an alternative estimation approach.

6.4.2. Asymmetric preferences

As we discussed at the start of Section 6, in many cases the econometrician is faced with several modeling alternatives when attempting to rationalize a given distribution of observables. So far, we have assumed that all bidders had the same preferences (either risk averse or risk neutral), but we have allowed distributions of valuations to vary across bidders. This allows us to reconcile bid distributions that vary across bidders. However, a natural alternative is that the distribution of valuations is the same for all bidders, but preferences differ. As mentioned above, in an ascending auction, behavior depends only on a bidder's valuation for the object, so it is impossible to distinguish these two cases. In a first-price auction with a fixed number of bidders, it is also difficult to distinguish these cases: in particular, as long as $\frac{G_{M_i|B_i}(b_i|b_i)}{g_{M_i|B_i}(b_i|b_i)}$ is increasing for each i, it is possible to rationalize bidding data using a model with homogeneous preferences. More generally, following the logic outlined above, there will generally exist homogeneous preferences with sufficient risk aversion such that

$$b_i + \lambda^{-1}\left(\frac{1}{n-1}\frac{G_{M_i|B_i}(b_i|b_i)}{g_{M_i|B_i}(b_i|b_i)}\right)$$

is increasing for all i.

However, there may be settings in which institutional information leads the econometrician to believe that a model with heterogeneous preferences is more natural than a model with heterogeneous value distributions. Campo (2002) has recently explored a model in which different bidders are permitted to have different preferences even though they draw their valuations from the same distribution. For $\alpha \in [0, 1]$ let $u_{i,\alpha}$ and $b_{i,\alpha}$ denote the αth quantile of bidder i's valuation and bid distributions, respectively. Generalizing our notation from above to allow bidders to have heterogeneous preferences represented by $\omega_i(\cdot)$, let $\lambda_i(s) \equiv \omega_i(s)/\omega_i'(s)$. Then, for all i, j, α, we have

$$u_{i,\alpha} = b_{i,\alpha} + \lambda_i^{-1}\left(\frac{G_{M_i}(b_{i,\alpha})}{g_{M_i}(b_{i,\alpha})}\right),$$

$$u_{j,\alpha} = b_{j,\alpha} + \lambda_j^{-1}\left(\frac{G_{M_j}(b_{j,\alpha})}{g_{M_j}(b_{j,\alpha})}\right).$$

Since the distributions of valuations are assumed to be the same across bidders, it follows that

$$b_{i,\alpha} + \lambda_i^{-1}\left(\frac{G_{M_i}(b_{i,\alpha})}{g_{M_i}(b_{i,\alpha})}\right) = b_{j,\alpha} + \lambda_j^{-1}\left(\frac{G_{M_j}(b_{j,\alpha})}{g_{M_j}(b_{j,\alpha})}\right). \tag{6.31}$$

Campo (2002) shows that a set of observed bid distributions that are independent and satisfy standard regularity conditions can be rationalized using this model if and only if (i) there exist functions $\lambda_1(\cdot), \ldots, \lambda_n(\cdot)$ such that (6.31) holds for every quantile $\alpha \in [0, 1]$ where, for each i, $\lambda_i(0) = 0$, $\lambda_i'(\cdot) \geqslant 1$, and (ii) $\xi_i(b) = b + \lambda_i^{-1}(\frac{G_{M_i}(b)}{g_{M_i}(b)})$ is strictly increasing. There is no guarantee that these conditions can be satisfied, because (6.31)

must hold for all $\alpha \in [0, 1]$.[80] Indeed, Campo (2002) provides an example where the conditions cannot be satisfied, and establishes that the set of bid distributions that can be rationalized using her model is a strict subset of those that can be rationalized using the model with homogeneous preferences and heterogeneous distributions of valuations.

Rather than analyze conditions under which risk preferences are nonparametrically identified, Campo (2002) takes a semi-parametric approach, with preferences given by $\omega(\cdot;\theta_i)$ (implying an associated $\lambda(\cdot;\theta_i)$), where θ_i is a finite dimensional parameter. To analyze identification, observe that for all i, j, α, α', we have

$$
\begin{aligned}
u_{i,\alpha} &= b_{i,\alpha} + \lambda^{-1}\left(\frac{G_{M_i}(b_{i,\alpha})}{g_{M_i}(b_{i,\alpha})};\theta_i\right),\\
u_{j,\alpha} &= b_{j,\alpha} + \lambda^{-1}\left(\frac{G_{M_j}(b_{j,\alpha})}{g_{M_j}(b_{j,\alpha})};\theta_j\right),\\
u_{i,\alpha'} &= b_{i,\alpha'} + \lambda^{-1}\left(\frac{G_{M_i}(b_{i,\alpha'})}{g_{M_i}(b_{i,\alpha'})};\theta_i\right),\\
u_{j,\alpha'} &= b_{j,\alpha'} + \lambda^{-1}\left(\frac{G_{M_j}(b_{j,\alpha'})}{g_{M_j}(b_{j,\alpha'})};\theta_j\right).
\end{aligned}
$$

Suppose, for example, that each θ_i is a scalar. Since by assumption $u_{i,\alpha} = u_{j,\alpha}$ and $u_{i,\alpha'} = u_{j,\alpha'}$, for a given pair of quantiles α and α', this is a system of four equations in four unknowns $(u_{i,\alpha}, u_{i,\alpha'}, \theta_i, \theta_j)$, so that θ_i and θ_j are identified using data from just two quantiles. Once θ_i and θ_j are known, $F_U(\cdot)$ is uniquely determined by the first-order conditions and the observed distribution $G_{M_i}(\cdot)$. Similarly, once $F_U(\cdot)$ is identified, the first-order conditions and $G_{M_k}(\cdot)$ determine θ_k for $k \neq i, j$. Campo (2002) considers the case of CRRA preferences discussed above and gives the nonsingularity conditions for the system of equations above (restrictions on the pair $(G_{M_i}(\cdot), G_{M_j}(\cdot))$) that ensure identification in that case. It is crucial that there are some asymmetries in the bid distributions. She proposes a parametric estimation approach. We refer readers to her paper for details.

Since it is possible to identify θ_i and θ_j using data from just two quantiles of the bidding distribution when θ_i is a scalar, there is no guarantee that, given observed bid distributions, a particular functional form can rationalize the data at every quantile. Indeed, the example considered by Campo (2002) of CRRA preferences $\omega_i(u) = u^{1-c_i}$ requires the existence of constants c_i, c_j on $[0, 1)$ such that

$$
b_{i,\alpha} + (1-c_i)\left(\frac{G_{M_i}(b_{i,\alpha})}{g_{M_i}(b_{i,\alpha})}\right) = b_{j,\alpha} + (1-c_j)\left(\frac{G_{M_j}(b_{j,\alpha})}{g_{M_j}(b_{j,\alpha})}\right). \tag{6.32}
$$

[80] Campo (2002) requires the condition $\lambda_i'(\cdot) \geqslant 1$ in order to guarantee that the induced preferences satisfy risk aversion; in fact, existence of an equilibrium in increasing strategies requires a slightly weaker condition, namely that $\ln(\omega_i(\cdot))$ is concave in the relevant region, which is guaranteed if $\lambda_i'(\cdot) \geqslant 0$. To see why log-concavity of $\omega_i(\cdot)$ is important, note that in an IPV auction, a bidder's objective function is $\omega_i(b_i - u_i)G_{M_i}(b_i)$. Maximizing this is equivalent to maximizing its logarithm; but, if $\ln(\omega_i(\cdot))$ is strictly convex, then $\frac{\partial^2}{\partial b_i \partial u_i}\omega_i(b_i - u_i) < 0$, so that bidders with higher valuations choose lower bids.

Suppose, for example, that there are just two bidders and that B_1 is uniformly distributed on [0, 1]. Then, (6.32) becomes, for all $\alpha \in [0, 1]$,

$$\alpha + (1 - c_1)\left(\frac{G_{B_2}(\alpha)}{g_{B_2}(\alpha)}\right) = (2 - c_2)G_{B_2}^{-1}(\alpha). \tag{6.33}$$

Clearly, this places strong restrictions on $G_{B_2}(\cdot)$. For example, this would rule out a distribution of the form $G_{B_2}(b_2) = b_2^{\gamma}$ with $\text{supp}[B_2] = [0, 1]$ (unless $\gamma = 1$, which would violate the assumption of asymmetric bid distributions).

Finally, we note that even when the data can be rationalized by both the homogeneous preference-heterogeneous valuations and the heterogeneous preference-homogeneous valuations models, it may be possible to extend the testing approaches described above that exploit exogenous variation in participation or other exclusion restrictions. When each model is identified for fixed $\mathcal{N}$, exogenous variation in $\mathcal{N}$ leads to over-identifying restrictions. In general, even if two different models rationalize the same data for fixed $\mathcal{N}$, the out-of-sample predictions of the models for $\mathcal{N}' \neq \mathcal{N}$ will differ between the two models. When data from auctions with both $\mathcal{N}$ and $\mathcal{N}'$ are observed, the out-of-sample predictions might be tested.

7. Common values auctions

While we have discussed a wide range of private values models in the preceding sections, in many applications a common values model may seem more natural. Recall that we use the term "common values" to refer to a broad class of models in which information about each bidder's valuation is dispersed among bidders (see Section 2).We emphasize, however, that the presence of factors affecting all bidders' valuations need not imply common values. For example, if

$$X_i = U_i = V_0 + \varepsilon_i$$

this is a private values specification despite the "common" factor V_0. Indeed, in this example each bidder knows his own valuation with certainty.[81] The presence of V_0 does introduce correlation of bidders' valuations and of bidders' information, and even causes one bidder's signal to be correlated with another's valuation; however, it does not introduce common values because no opponent has information that is relevant to a bidder's assessment of his own valuation, given that he has observed his own signal. The critical distinction concerns the nature of bidders' private information. When each bidder's private information concerns only idiosyncratic determinants of his own valuation, this is a private values setting.

[81] This is not essential for a private values environment. For example, if $U_i = X_i + \epsilon_i$ with ϵ_i independent of X_j for all $j \neq i$, this remains a private values setting.

Nonetheless, many auction environments seem likely to fall in the common values category. Often the good for sale will not be consumed immediately (or the procurement contract being bid for will not be fulfilled immediately), and bidders may have different information about future states of the world – e.g., market conditions or the supply and demand of substitute objects. In some applications bidders will naturally have access to different information. A bidder might conduct her own seismic survey of an oil tract or might learn about market conditions from her own customers and suppliers. Furthermore, even if bidders have access to the same market data, they may have different algorithms or rules-of-thumb for using this information to form beliefs about the object's value. The output of one bidder's algorithm (i.e., its signal) might then be useful to another bidder in assessing her own valuation even after seeing the output of her own algorithm. In such cases it may be appropriate to model bidders as having different private information of a common values nature.

Aside from the potential prevalence of common values in practice, common values models are also of particular interest because they provide an example of a market environment in which adverse selection may play an important role. In a private values auction, bidders need only to follow a simple dominant ("bid your value") strategy in an ascending auction or to respond optimally to a distribution of opposing bids in a first-price auction. In a common values auction, bidders must understand the strategies that underlie the competing bids in order to make correct inferences about their informational content; in particular, bidders must account for the information that would be implied by their winning the auction in order to avoid the winner's curse. An important contribution of the empirical industrial organization literature has been to confirm some of the fairly subtle equilibrium predictions of common values auction models.[82] However, a number of positive and normative questions depend not just on whether bidder behavior is broadly consistent with theory, but on the exact structure of demand and information.

For example, typically the seller or auctioneer has some discretion over the auction rules. As first demonstrated by Milgrom and Weber (1982) for symmetric common values environments, the information revealed publicly by losing bidders' exits in an ascending auction reduces both the severity of the winner's curse and the informational rents obtained by the winner, leading to higher expected revenues than with a first-price sealed-bid auction. With asymmetries, first-price auctions may allocate the good inefficiently; however, they tend to raise more revenue in private values settings, may be less susceptible to collusion (detection and response to defections are more difficult than in an ascending auction), and may be less costly to administer. The choice of auction format also affects bidder entry when bidders are asymmetric [Klemperer (2002), Athey, Levin and Seira (2004)]. In trading off these factors, a seller must understand the underlying structure of bidder demand and information that determines the significance

[82] Examples include Hendricks, Porter and Boudreau (1987), Hendricks and Porter (1988), Hendricks, Porter and Wilson (1994), Athey and Levin (2001), and Haile (2001).

of each factor. Even within an auction format, the joint distribution of signals and valuations is important for positive questions (e.g., the division of surplus) and for design issues (e.g., the optimal reserve price, the optimal entry fee, and whether restrictions on participation would be profitable).[83]

7.1. Limits of identification with a fixed number of bidders

In a common values environment, identifying the joint distribution $F_{\mathbf{X},\mathbf{U}}(\cdot)$ requires substantial restrictions on the underlying structure, and/or data beyond bids from a fixed environment. To suggest why, observe that in common values auctions the primitives of the model involve two different random variables for each bidder i: X_i and U_i. Hence, the joint distribution $F_{\mathbf{X},\mathbf{U}}(\cdot)$ governs $2n$ random variables, yet an auction will reveal at most n bids.[84] Even in the special case of pure common values, where $U_i = U_0$ for all i, the primitive of interest, $F_{\mathbf{X},U_0}(\cdot)$ has dimension $n+1$. So some additional structure will be necessary to obtain identification.

We begin by considering first-price auctions. One convenient normalization of signals (recall that this is without loss of generality) is[85]

$$E\Big[U_i \;\Big|\; X_i = \max_{j\neq i} X_j = x, \mathcal{N}\Big] = x. \tag{7.1}$$

With this normalization, (2.3) and the first-order condition (2.4) imply

$$v_i(x_i, x_i; \mathcal{N}) = x_i = b_i + \frac{G_{M_i|B_i}(b_i|b_i; \mathcal{N})}{g_{M_i|B_i}(b_i|b_i; \mathcal{N})}. \tag{7.2}$$

In Section 3.2 we discussed the identification and estimation of the distribution of the random variable

$$B_i + \frac{G_{M_i|B_i}(B_i|B_i; \mathcal{N})}{g_{M_i|B_i}(B_i|B_i; \mathcal{N})}.$$

All that changes when we consider common values settings is the interpretation of this distribution: using (7.2) we now interpret it as the distribution of the random variable $v_i(X_i, X_i; \mathcal{N})$.

This distribution alone will be sufficient for some questions of interest (see, e.g., Section 8), but certainly not all. In particular, it does not provide identification of the joint distribution $F_{\mathbf{U},\mathbf{X}}(\cdot)$. Consider the case in which one observes all bids from auctions with n symmetric bidders. Under the private values assumption, $v(X_i, X_i; n) = X_i = U_i$,

[83] In a common values auction, restricting participation reduces the severity of the winner's curse, leading to more aggressive bidding. This can result in higher expected revenues despite the presence of fewer bidders, depending on the underlying distributions [e.g., Smiley (1979), Matthews (1984), Hong and Shum (2002)].

[84] Note that a normalization of signals does not change this argument, since the normalization cannot address the correlation between signals and valuations.

[85] Note, however, that this normalization cannot be maintained if $\mathcal{N}$ varies.

and the joint distribution $F_{\mathbf{U}}(\cdot)$ is just identified (see Section 3). Under the common values assumption the joint distribution of $(v(X_1, X_1; n), \ldots, v(X_n, X_n; n))$ is $F_{\mathbf{X}}(\cdot)$ under the normalization (7.1). Since this distribution is just identified when n is fixed, it follows that it is impossible to distinguish common values from private values based on bidding data from first-price auctions with no reserve price and a fixed number of bidders [Laffont and Vuong (1996), Guerre, Perrigne and Vuong (2000)]. Thus, it is important to emphasize that any conclusions from data from a fixed set of bidders (and with no reserve price) rely on a maintained assumption of common values or private values. For example, it might be possible to justify a wide range of reserve prices as optimal for the seller under different assumptions about $F_{\mathbf{U},\mathbf{X}}(\cdot)$ that are consistent with the identified marginal distribution $F_{\mathbf{X}}(\cdot)$.

Ascending auctions are even more difficult in the common values setting. First, just as in a first-price auction, it would be impossible to distinguish common values from private values using a data set with a fixed number of bidders, even if all bids (including the planned exit price of the winner) were observed. Any observed distribution of bids could simply be equal to the distribution of private values for the bidders [Laffont and Vuong (1996)]. Second, exactly as in the case of a private values ascending auction, the unobservability of the winner's planned exit price can challenge even the identification of $F_{\mathbf{X}}(\cdot)$. Further, while a normalization like (7.1) can be applied to signals in the *initial* phase of an ascending auction (the period before any bidders drop out), no single normalization can induce the simple strategy $\beta_i(x_i, n) = x_i$ throughout the auction, since bidders modify their strategies each time an opponent exits. The exact forms of these modifications depend on the joint distribution of signals and valuations. While we might hope that this dependence would enable observed bids to provide information about this joint distribution, it also creates serious challenges. Finally, further complications arise from the fact that, when $n > 2$, there is a multiplicity of symmetric equilibria in weakly undominated strategies in common values auctions [Bikhchandani, Haile and Riley (2002)], implying that there is no unique interpretation of bids below the transaction price.

The following result from Athey and Haile (2002) establishes that the common values model is generally not identified in ascending auctions. Here we ignore the multiplicity of equilibria and assume a special case of a pure common values model in which signals are i.i.d. Even this very restrictive common values model is not identified.

THEOREM 7.1. *In an ascending auction, assume the pure common values model, i.i.d. signals X_i, and select the equilibrium characterized by* Milgrom and Weber (1982). *With n fixed, the model is not identified* (*even up to a normalization of signals*) *from the observable bids.*

PROOF. Take $n = 3$ and consider two models. In both, signals are uniform on $[0, 1]$. In the first, the value of the good is

$$u_0 = u(x_1, x_2, x_3) = \frac{\sum_i x_i}{3},$$

while in the second model

$$u_0 = \hat{u}(x_1, x_2, x_3) = \frac{x^{(1:3)}}{3} + \frac{x^{(2:3)}}{6} + \frac{x^{(3:3)}}{2}.$$

Because in both models $E[U_0 \mid X_1 = X_2 = X_3 = x] = x$, equilibrium bidding in the initial phase of the auction is identical in the two models in the Milgrom–Weber equilibrium (see Section 2.2.2); i.e., $G_B^{(1:3)}(b) = F_X^{(1:3)}(b) = 1-(1-b)^3$ in both cases. Similarly, since $b^{(2:3)} = E[U_0 \mid X^{(1:3)} = b^{(1:3)}, X^{(3:3)} = X^{(2:3)} = x^{(2:3)}]$, the fact that $\hat{u}(x, y, y) = u(x, y, y)$ for all x and y implies that $G_B^{(2:3)}(\cdot|B^{(1:3)})$ is identical under the two models. Since $G_B^{(1:3)}(\cdot)$ and $G_B^{(2:3)}(\cdot|B^{(1:3)})$ completely determine the joint distribution of the observable bids, the two models are observationally equivalent. □

This is a strong negative result for common values ascending auctions. Even ignoring the equilibrium selection problem and possible doubts about the interpretation of losing bids in an ascending auction, this most restrictive of common values models is not identified. This nonidentification is important for policy. Continuing the example from the proof of Theorem 7.1, consider the simple problem of setting an optimal reserve price for a second-price sealed-bid auction. Recalling the participation threshold (2.1), the optimal reserve price solves

$$\max_r 3(1 - F_X)\big(x^*(r, 3)\big) F_X\big(x^*(r, 3)\big)^2 r + \int_{x^*(r,3)}^1 v(y, y; 3) 6 f_X(y)\big(1 - F_X(y)\big) F_X(y)\, \mathrm{d}y.$$

By construction, $v(x, x; 3)$ is the same for all x in the two models. However, for any r, the participation threshold $x^*(r, 3)$ is lower in the second model, due to the reduced dependence of each U_i on X_{-i} when X_i is maximal. Hence, the objective function above differs for the two models and (as can be confirmed directly) implies different optimal reserve prices.

We will see in Section 8 that variation in the number of bidders can be useful for overcoming at least some of these limitations. In particular, this variation can be sufficient to enable discrimination between private and common values models. Whether this kind of variation can go farther to enable nonparametric identification of a common values model is a question not yet explored. Below we will consider identification through additional structure and through additional data.[86]

7.2. Pure common values

Given the negative identification results obtained thus far, it is natural to consider whether additional assumptions can alleviate the problem. One possible approach is to restrict attention to the *pure* common values model, where $U_i = U_0 \; \forall i$.

[86] Parametric models of common values auctions have been estimated by, e.g., Smiley (1979), Paarsch (1992a), Hong and Shum (2002, 2003), and Bajari and Hortaçsu (2003a).

In the pure common values model, the joint distribution $F_{\mathbf{X},\mathbf{U}}(\cdot)$ governs $n+1$ random variables $(U_0, X_1, \ldots, X_n)$; however, at most n bids are revealed in a first-price auction, and only $n-1$ "bids" are revealed in the standard model of an ascending auction. This suggests that the pure common values assumption alone will not be sufficient to obtain identification, and that either additional structure or additional data will be needed. Below, we explore examples of both: we first consider additional restrictions on $F_{\mathbf{X},U_0}(\cdot)$, and then consider cases in which the realization of U_0 is observable *ex post*.

7.2.1. Identification with additional structure: The mineral rights model

A special case of the pure common values model given considerable attention in the literature is the symmetric "mineral rights model" defined in Section 2.1. Here, bidders' signals are i.i.d. conditional on the realization of the common value U_0. As the name suggests, this model is motivated by auctions in which firms bid for the right to extract oil from an offshore tract. All firms may place the same value on the oil, since it is sold in a common market, but none knows how much oil (if any) there is. Each receives a seismologist's report, providing a (conditionally independent) noisy signal of U_0. This structure may be natural in other applications as well.

Even with this structure, identification from bid data is not straightforward since this requires somehow separating the variation in bids due to the randomness of U_0 from that due to the randomness of X_i conditional on u_0. One possible approach is to assume a separable functional form like $X_i = U_0 + A_i$, where the "errors" A_i mutually independent conditional on U_0. This can be useful, although the additive structure need not survive the normalization (7.1) in general. Put differently, while it will be useful for the left-hand side of the first-order condition (7.2) to have a separable form, one must be careful about what underlying structures on (X_i, U_0) can deliver this separability. This is a question that has been explored by Li, Perrigne and Vuong (2000). To discuss their approach, we first define two (nested) special cases of the symmetric mineral rights model.

Linear Mineral Rights (LMR): $U_i = U_0$. In addition, for each n there exist two known constants $(C, D) \in \mathbb{R} \times \mathbb{R}_+$ and random variables $(A_1, \ldots, A_n)$ with joint distribution $F_{\mathbf{A}}(\cdot)$ such that, with the normalization $E[U_0 \mid X_i = \max_{j \neq i} X_j = x, n] = x$, either (i) $X_i = \exp(C) \cdot (U_0 \cdot A_i)^D \ \forall i$, with (A_i, U_0, X_i) nonnegative; or (ii) $X_i = C + D(U_0 + A_i) \ \forall i$. Further, conditional on U_0, the components of $\mathbf{A}$ are mutually independent and identically distributed.

LMR with Independent Components (LMR-I): In the LMR model, $(U_0, \mathbf{A})$ are mutually independent, with all A_i identically distributed.

Li, Perrigne and Vuong (2000) focus on the LMR-I model and provide examples satisfying its assumptions. Under the LMR-I model, taking case (ii), (7.2) simplifies to

$$C + D(u_0 + a_i) = b_i + \frac{G_{M_i|B_i}(b_i|b_i; n)}{g_{M_i|B_i}(b_i|b_i; n)}. \tag{7.3}$$

Since C and D are known and the right-hand side of (7.3) is observable, it follows that the joint distribution of $(U_0 + A_1, \ldots, U_0 + A_n)$ is identified from a data set containing all bids in first-price auctions. Li, Perrigne and Vuong (2000) note that standard deconvolution results, such as those used in the literature on measurement error (see Section 6.1.2), can then be used to separately identify the distributions $F_{U_0}(\cdot)$ and $F_A(\cdot)$.[87]

THEOREM 7.2. *Assume that for all i, the characteristic functions $\psi_{U_0}(\cdot)$ and $\psi_{A_i}(\cdot)$ of the random variables U_0 and A_i are nonvanishing. If all bids are observed in a first-price auction, then the LMR-I model is identified.*

Even with these kinds of strong assumptions, identification is problematic when some bids are unobserved. Bids reveal realizations of order statistics of the form $U_0 + A^{(i:n)}$. Since order statistics are correlated even when the underlying random variables are independent, the identification approach based on the measurement error literature followed by Li, Perrigne and Vuong (2000) fails, unless all order statistics are observed (impossible in an ascending auction).

7.2.2. Identification and testing when ex post values are observable

In some applications, an *ex post* measure of the realized common value u_0 will be observable to the econometrician. One notable example is an US outer-continental-shelf auction of drilling rights, where the quantities of oil and other minerals extracted from a tract are metered [e.g., Hendricks and Porter (1988), Hendricks, Pinkse and Porter (2003)]. Another example is a "scaled sale" timber auction, common in the US and Canada, where the quantity of each species of timber extracted from a tract is recorded by an independent agent at the time of harvest [e.g., Athey and Levin (2001)]. In other cases, resale prices can provide measures of realized values [e.g., McAfee, Takacs and Vincent (1999)]. Such additional data can be helpful in the mineral rights model.[88] In practice, the measures of u_0 available may be only imperfectly correlated with the true value to the bidders; we discuss this possibility below.

7.2.2.1. First-price auctions When we impose the structure of the symmetric pure common values model, the first-order condition (7.2) and the normalization (7.1) give

$$E\Big[U_0 \;\Big|\; X_i = \max_{j\neq i} X_j = x_i, n\Big] = x_i = b_i + \frac{G_{M_i|B_i}(b_i|b_i;n)}{g_{M_i|B_i}(b_i|b_i;n)}. \tag{7.4}$$

[87] Février (2004) has recently proposed an interesting alternative restriction of the mineral rights model that enables identification. He considers the case in which, conditional on u_0, each X_i has support $[\underline{u}, u_0]$ and density $f_{x|u_0}(\cdot) = \frac{h(\cdot)}{H(u_0)}$ for some function $H(\cdot)$ with derivative $h(\cdot)$ satisfying $H(\underline{u}) = 0$. With this structure, conditional on having the highest signal, there is no information in the signals of one's opponents. He shows that this structure enables identification up to scale.

[88] Smiley (1979) was the first to suggest the value of such information. In his application he did not have access to an *ex post* measure and instead explored use of a noisy *ex ante* measure.

When all bids and the realization of U_0 are observed, (7.4) enables identification of the joint distribution $F_{\mathbf{X},U_0}(\cdot)$.

With knowledge of $F_{\mathbf{X},U_0}(\cdot)$, it is possible to perform counterfactual experiments, quantify the extent to which information is dispersed among the bidders, and characterize the magnitude of the "winner's curse." For example, it is interesting to examine the differences

$$E[U_0 \mid X_i = x_i, n] - E\Big[U_0 \;\Big|\; X_i = \max_{j\neq i} X_j = x_i, n\Big] \tag{7.5}$$

and

$$E[U_0 \mid X_i = x_i, n] - E\Big[U_0 \;\Big|\; X_i = x_i,\ \max_{j\neq i} X_j \leqslant x_i, n\Big] \tag{7.6}$$

since these provide a measure bidders' equilibrium responses to the winner's curse under the pure common values assumption.[89]

Hendricks, Pinkse and Porter (2003) were the first to suggest this and also proposed a test of equilibrium bidding in this model. Let

$$\zeta(b_i, b_j, n) = E\Big[U_0 \;\Big|\; B_i = b_i,\ \max_{j\neq i} B_j = b_j, n\Big]. \tag{7.7}$$

When the equilibrium bid function $\beta(\cdot; n)$ is strictly increasing, $\beta(x_i; n) = b_i$ implies

$$\begin{aligned}
&E\Big[U_0 \;\Big|\; X_i = \max_{j\neq i} X_j = x_i, n\Big] \\
&\quad = E\Big[U_0 \;\Big|\; \beta(X_i; n) = \max_{j\neq i} \beta(X_j; n) = \beta(x_i; n), n\Big] \\
&\quad = E\Big[U_0 \;\Big|\; B_i = \max_{j\neq i} B_j = b_i, n\Big] \\
&\quad = \zeta(b_i, b_i, n).
\end{aligned}$$

Thus, the first-order condition (7.4) can be written

$$\zeta(b_i, b_i, n) = b_i + \frac{G_{M|B}(b_i|b_i; n)}{g_{M|B}(b_i|b_i; n)} \equiv \xi(b_i, n). \tag{7.8}$$

[89] Hendricks, Pinkse and Porter (2003) point out that a positive value for the difference in (7.6) cannot be used as evidence against a private values assumption. The problem is that the interpretation of the empirical measure of u_0 as the realized value of the good relies on the pure common values assumption. For example, consider a symmetric independent private values environment and suppose the measured "*ex post* value" u_0 is actually just $\max_j u_j$, i.e., the value to the winner. Then the difference (7.6) is

$$E\Big[\max_j U_j \;\Big|\; U_i = u_i, n\Big] - E\Big[\max_j U_j \;\Big|\; U_i = u_i,\ \max_{k\neq i} U_k \leqslant u_i, n\Big]$$

which is positive. We will discuss approaches that can be used to discriminate private from common values models in Section 8.

Note that because the joint distribution of $(U_0, \mathbf{B}, N)$ is observable, $\zeta(b_i, b_i, n)$ is identified directly through Equation (7.7). No behavioral assumption is required for this identification: $\zeta(b_i, b_i, n)$ is simply a conditional expectation of the observable U_0 given that the observable bids satisfy $B_i = \max_{j \neq i} B_j = b_i$. Since $\xi(b_i, n)$ is also identified from the bidding data under the assumption of equilibrium bidding, the overidentifying restriction $\zeta(b_i, b_i, n) = \xi(b_i, n)$ can be tested.

To examine the differences (7.5) and (7.6) empirically, Hendricks, Pinkse and Porter (2003) first observe that since $b_i = \beta_i(x_i; n)$ and bidding is strictly monotonic, these differences are equal to the differences

$$E[U_0 \mid B_i = b_i, n] - E\Big[U_0 \;\Big|\; B_i = \max_{j \neq i} B_j = b_i, n\Big]$$

and

$$E[U_0 \mid B_i = b_i, n] - E\Big[U_0 \;\Big|\; B_i = b_i,\ \max_{j \neq i} B_j \leqslant b_i, n\Big].$$

They suggest a univariate local linear estimator $\hat{w}(b; n)$ of $E[U_0 \mid B_i = b, n]$, where $\hat{w}(b; n)$ is the solution for w in the problem

$$\min_{w,\gamma} \sum_{t=1}^{T_n} \frac{1}{n} \sum_{i=1}^{n} \big(u_0 - w - \gamma(b - b_{it})\big)^2 \mathbf{1}\{n_t = n\} K\left(\frac{b - b_{it}}{h}\right)$$

with $K(\cdot)$ denoting a kernel and h a bandwidth [see, e.g., Loader (1999)]. A similar estimator for $E[U_0 \mid B_i = \max_{j \neq i} B_j \leqslant b_i, n]$ is obtained by using only the winning bid b_i from each auction t, rather than all bids.

To examine the overidentifying restriction (7.8), the right-hand side can be estimated using the kernel methods described in Section 3.2.1. A bivariate local linear estimator $\hat{v}(b)$ of the conditional expectation $\zeta(b_i, b_i, n) = E[U_0 \mid B_i = \max_{j \neq i} B_j = b_i, n]$ is obtained from the solution to

$$\min_{v,\gamma_1,\gamma_2} \sum_{t=1}^{T_n} \frac{1}{n} \sum_{i=1}^{n} \big(u_0 - v - \gamma_1(b - b_{it}) - \gamma_2(b - m_{it})\big)^2 \mathbf{1}\{n_t = n\} \times K\left(\frac{b - b_{it}}{h_1}\right) K\left(\frac{b - m_{it}}{h_2}\right).$$

Here m_{it} is the maximum realized bid among i's opponents at auction t. Hendricks, Pinkse and Porter (2003) suggest the use of the bootstrap to construct confidence intervals for testing.

Because of their focus on testing, Hendricks, Pinkse and Porter (2003) did not explore estimation of the joint distribution $F_{\mathbf{X},U_0}(\cdot)$. However, with the normalization (7.1), an estimate of

$$E\Big[U_0 \;\Big|\; B_i = b_i,\ \max_{j \neq i} B_j = b_i, n\Big]$$

(such as that obtained using the first-order condition and kernel methods described in Section 3.2.1) provides an estimate of each realized x_i. Combining this with the observable realizations of U_0 presumably would enable consistent estimation of the joint distribution $F_{\mathbf{X},U_0}(\cdot)$ and/or the associated density $f_{\mathbf{X},U_0}(\cdot)$.

7.2.2.2. Ascending auctions Since the common values model is over-identified in a first-price auction when *ex post* values are observed, one might hope for identification in an ascending auction with similar data. However, the partial observability of bids in ascending auctions again presents serious challenges. Consider the case of two symmetric bidders. Recall that in an ascending auction, $\beta_i(X_i; n)$ is bidder i's planned exit price when his opponent has not yet exited the auction. With $n = 2$ there is no problem of multiple equilibria [Bikhchandani, Haile and Riley (2002)], and

$$\begin{aligned} b_i &= \beta_i(x_i; 2) = E[U_0 \mid X_1 = X_2 = x_i,\ n = 2] \\ &= E\big[U_0 \mid \beta_1(X_1; 2) = \beta_2(X_2; 2) = \beta_i(x_i; 2),\ n = 2\big] \\ &\equiv \zeta(b_i, b_i, 2). \end{aligned}$$

However, it is not possible to estimate

$$\zeta(b_1, b_2, 2) = E[U_0 \mid B_1 = b_1,\ B_2 = b_2]$$

directly from the data (as is possible in a first-price auction) since in any given auction we observe the exit price of only one bidder. We never observe B_1 and B_2 in the same auction. We can observe the joint distribution of $(U_0, B^{(1:2)})$. Under (7.1), this is also the distribution of $(U_0, X^{(1:2)})$, but without additional structure this information is not sufficient to recover $F_{U_0,\mathbf{X}}(\cdot)$.

If we impose the additional structure of the mineral rights model, however, then conditional on U_0, $X^{(1:2)}$ is an order statistic from a sample of independent draws from $F_{X|U_0}(\cdot)$. Exploiting Equation (4.1), identification is then obtained in the symmetric two bidder case when the transaction price and *ex post* value are observed. Extending this approach to the case with $n > 2$ symmetric bidders is possible as well, using the order statistic $B^{(1:n)}$, although the suitability of this extension may be doubted in practice. To see one possible approach, note that with the normalization

$$x_i = E\big[U_0 \mid X_j = x_i\ \forall j \in \{1, \ldots, n\}\big]$$

we have $B^{(1:n)} = X^{(1:n)}$. Exploiting (4.1) again, we could recover the distribution of $X_i|U_0$, which then delivers identification of $F_{U_0,\mathbf{X}}(\cdot)$. This relies on the interpretation of bids implied by the button auction model, which may be especially dubious when applied to the interpretation of the lowest bid. To apply a similar approach using the transaction price $B^{(n-1:n)}$, it is still necessary to make use of the losing bids because the bidders themselves condition on this information. However, it may be easier to defend an approach which incorporates the information from all bids than one based entirely on the lowest bid. Here, we sketch one possibility.

Fix a set of realized values for the $n-2$ lowest bids at $(b^{(1:n)}, \ldots, b^{(n-2:n)})$. Then, the (observable) distribution of

$$B^{(n-1:n)} \mid \left\{U_0 = u_0,\, B^{(1:n)} = b^{(1:n)}, \ldots, B^{(n-2:n)} = b^{(n-2:n)}\right\}$$

is equal to the distribution of

$$B^{(n-1:n)} \mid \left\{U_0 = u_0,\ B^{(n-1:n)} \geqslant b^{(n-2:n)}\right\} \tag{7.9}$$

since bids (which are strictly increasing functions of the signals) are independent conditional on U_0 in the mineral rights model. Consider the following normalization of signals:

$$x = E\left[U_0 \mid X_{n-1} = X_n = x,\ B^{(1:n)} = \cdots = B^{(n-2:n)} = \inf\left[\operatorname{supp}[B_i]\right]\right].$$

Imposing this normalization, when $b^{(n-2:n)} = \inf[\operatorname{supp}[B_i]]$, the random variable in (7.9) is equal in distribution to

$$X^{(n-1:n)}|U_0 = u_0.$$

Equation (4.1) then uniquely determines the distribution of $X_i|U_0$ and thus $F_{U_0,\mathbf{X}}(\cdot)$, since U_0 is directly observable.

This result is not as strong as one might hope for. It relies on an interpretation of the losing bids in an ascending auction (although it is not essential that the bidders use Milgrom and Weber's (1982) equilibrium) and on an assumption that the econometrician's inferences about exit prices match those that the bidders make during the auction. Furthermore, the identification argument relies on the tails of the distribution of bids. In particular, building a nonparametric estimator based on this argument would seem to require an estimate of the distribution of $B^{(n-1:n)}$ conditional on both the *ex post* value u_0 and the event that $n-2$ losing bids are increasingly close to the bottom of the support of the bid distribution. Whether this kind of approach can work well in sample sizes typically available has not been investigated.

7.2.2.3. Biased or noisy observations of ex post values So far in this section we have assumed that the econometrician observes the true realization of U_0. In the case of an oil auction, for example, this requires that the oil extracted is measured without error and that the econometrician has accurate measures of all costs (including opportunity costs) incurred in extracting the oil. These may be strong assumptions in some applications, so it is useful to consider the degree to which they can be relaxed. With the exception of Yin (2004), which does not fully address identification, the literature has not analyzed the issue of imperfect measures of *ex post* values.[90] Here, we present some initial results.

[90] Yin (2004) obtained descriptions of items auctioned on eBay and recruited volunteers to make subjective assessments of the value of the objects. The mean assessment was then treated as a potentially biased proxy for U_0. Smiley (1979, Appendix), explored the use of a noisy *ex ante* measure of U_0 within a parametric model.

Consider a first-price auction in which all bids are observable and suppose the available measure of U_0 is

$$\widetilde{U}_0 = \gamma_0 + \gamma_1 U_0 + \varepsilon, \tag{7.10}$$

where γ_0 and γ_1 are fixed parameters, unknown to the econometrician, and ε is an unobserved random variable satisfying $E[\varepsilon \mid \mathbf{X} = \mathbf{x}] = 0$ for all $\mathbf{x}$. Recall that, given the normalization (7.1), $F_{\mathbf{X}}(\cdot)$ is identified from the bidding data alone in this setting. Since we observe $\widetilde{U}_0$ in every auction, the distribution $F_{\mathbf{X},\widetilde{U}_0}(\cdot)$ is identified as well. With this, we can compute

$$\eta(x) \equiv E\Big[\widetilde{U}_0 \;\Big|\; X_i = \max_{j\neq i} X_j = x\Big].$$

Given (7.10), the normalization (7.1), and $E[\varepsilon \mid \mathbf{X} = \mathbf{x}] = 0$, we also have

$$\eta(x) = \gamma_0 + \gamma_1 E\Big[U_0 \;\Big|\; X_i = \max_{j\neq i} X_j = x\Big] = \gamma_0 + \gamma_1 x. \tag{7.11}$$

Since $\eta(\cdot)$ is identified, the joint distribution of $(\eta(X), X)$ is identified, as is the bias in the measure of the common value, determined by the parameters γ_0 and γ_1.

Identification of γ_0, γ_1, and $F_{\mathbf{X},\widetilde{U}_0}(\cdot)$ implies identification of quantities such as

$$E[U_0 \mid X_i = x] = E[\widetilde{U}_0 - \gamma_0 \mid X_i = x]/\gamma_1,$$

and

$$E\Big[U_0 \;\Big|\; X_i = x,\ \max_{j\neq i} X_j \leqslant x\Big] = E\Big[\widetilde{U}_0 - \gamma_0 \;\Big|\; X_i = x,\ \max_{j\neq i} X_j \leqslant x\Big]/\gamma_1,$$

so that the differences (7.5) and (7.6) discussed above are identified, for example. Unfortunately, unless ε is degenerate, the variance of U_0 is not identified, nor is the joint distribution $F_{\mathbf{X},U_0}(\cdot)$. In the setting studied by Yin (2004), where $\widetilde{U}_0$ is the mean estimate from a survey, the assumption that ε is degenerate may be a reasonable approximation when the number of survey respondents per auction is large.

In ascending auctions, the analysis is more complex. Let us focus on the case of a pure CV model with two bidders. Let us maintain the assumption $E[\varepsilon \mid \mathbf{X} = \mathbf{x}] = 0$ for all $\mathbf{x}$, and in addition, assume that $(X_1, \ldots, X_n)$ are independent conditional on $\widetilde{U}_0$. The joint distribution of $(B^{(1:2)}, \widetilde{U}_0)$ is then identified and, under (7.1), this is equal to the joint distribution of $(X^{(1:2)}, \widetilde{U}_0)$. Given that X_1 is independent of X_2 conditional on $\widetilde{U}_0$, the parent distribution $F_X(\cdot)$ is identified using (4.1), so that the joint distribution of $(\mathbf{X}, \widetilde{U}_0)$ is identified. This completely determines $\eta(\cdot)$, which in turn yields identification of γ_0 and γ_1 through (7.11).

Of course, the assumption that $\mathbf{X}$ is independent conditional on $\widetilde{U}_0$ may be strong, especially if $\widetilde{U}_0$ is a noisy measure of U_0. In practice this may be most defensible when ε is degenerate, so that $\widetilde{U}_0$ is a deterministic function of U_0.

8. Private versus common values: Testing

Negative identification results for common values models provide one important motivation for formal tests that could distinguish between common and private values models. Distinguishing private values from common values was, in fact, the goal behind Paarsch's (1992a) pioneering work on structural empirical approaches to auctions. The distinction between the two paradigms is central to our understanding of behavior in auction markets and has important implications for market design. For example, revenue superiority of an ascending auction relative to a second-price sealed-bid auction in symmetric settings [Milgrom and Weber (1982)] holds only in a common values environment. Furthermore, a common values environment is one with adverse selection. There is relatively little evidence on the empirical significance adverse selection, and an examination of the prevalence of common values in auctions might be suggestive of the nature of private information in other market environments as well.

It might be surprising that questions about the qualitative nature of private information could be answered at all empirically. In fact, early approaches to testing based on reduced-form relationships between bids and the number of bidders were eventually discovered to be invalid. With the structural approach proposed by Paarsch (1992a), it was possible to test particular common values or private values models, but only with maintained parametric distributional assumptions. More recently, Laffont and Vuong (1996) have pointed out that private values and common values models are empirically indistinguishable, suggesting that testing was impossible (see Section 7.1).[91] However, they did not consider the possibilities created by variation in the number of bidders or a binding reserve price. Below we will show how either of these can offer approaches for discriminating between private and common values.

In the case of variation in the number of bidders, the idea is simple. The winner's curse is present only in common values models and becomes more severe as more competitors are added. Having a signal of the object's value that is the highest among twenty implies a more severely biased signal than does having the highest signal among two, for example. This greater severity manifests itself as a reduction in a bidder's expectation of his valuation conditional on winning a large auction. In particular, while the unconditional expectation $E[U_i \mid X_i = x_i]$ is invariant to the set of opponents i faces, his equilibrium bid reflects a downward adjustment

$$E[U_i \mid X_i = x_i] - v_i(x_i, x_i; \mathcal{N})$$

that accounts for the information implied by his bid being pivotal. In a symmetric environment, this downward adjustment is always larger when i faces more competition.

To make this precise, suppose that the number of bidder is exogenous and let $\mathcal{N}_{+j}$ denote the set of bidders comprised of all members of $\mathcal{N}$ plus bidder j.[92]

[91] This can be thought of as a nonidentification result for the affiliated values model, which nests the private and common values models.

[92] Note that because the normalization (7.1) depends on the set of bidders, we could not maintain this particular normalization for all $\mathcal{N}$. Other normalizations, e.g., $x_i = F_{X_i}(x_i)$, are of course possible.

LEMMA 8.1. *Suppose the number of bidders varies exogenously. With private values, $v_i(x, x; \mathcal{N}) = v_i(x, x; \mathcal{N}_{+j})$ for all x, $\mathcal{N}$, i, j. With common values and symmetric bidders, $v_i(x, x; \mathcal{N}) > v_i(x, x; \mathcal{N}_{+j})$ for all x, $\mathcal{N}$, i, j.*

PROOF. The first claim is immediate from the fact that $v_i(x, x; \mathcal{N}) = x$ with private values. With common values and symmetric bidders,

$$v_i(x, x; \mathcal{N}) = E\Big[U_i \;\Big|\; X_i = x,\ \max_{k \in \mathcal{N}_{-i}} B_k = \beta(x; \mathcal{N})\Big].$$

Let $m = \arg\max_{k \in \mathcal{N}_{-i}} B_k$ and suppose $j \notin \mathcal{N}$. Then

$$\begin{aligned}
&v_i(x, x; \mathcal{N}) \\
&\quad = E\big[U_i \mid X_i = x,\ B_m = \beta_i(x; \mathcal{N}),\ B_k \leqslant \beta_i(x; \mathcal{N})\ \forall k \in \mathcal{N} \text{ s.t. } k \neq i, m\big] \\
&\quad = E_{X_j}\big[E\big[U_i \mid X_i = x,\ X_m = x,\ X_k \leqslant x\ \forall k \in \mathcal{N} \text{ s.t. } k \neq i, m\big]\big] \\
&\quad > E\big[U_i \mid X_i = x,\ \max\{X_m, X_j\} = x,\ X_k \leqslant x\ \forall k \in \mathcal{N} \text{ s.t. } k \neq i, j, m\big] \\
&\quad = v_i(x, x; \mathcal{N}_{+j})
\end{aligned}$$

where the last two lines follow from Definition 2.1 and the strict monotonicity of equilibrium bidding strategies.[93] □

This result provides the basis for testing using variation in the number of bidders. Although to our knowledge the proof was first given by Athey and Haile (2002) and Haile, Hong and Shum (2003), the idea behind this result and its potential value for detecting the winner's curse goes back at least to Gilley and Karels (1981), who suggested regressing bids from first-price auctions on the number of bidders as a test of a common values model. This reflected a belief that bids must increase with n in a private values auction (since adding bidders makes the auction more competitive) but might decline in n in a common values auction if the winner's curse were sufficiently severe to overcome the competitive effect of adding additional bidders [see, e.g., Brannman, Klein and Weiss (1987), Paarsch (1992a, 1992b), Laffont (1997)]. However, Pinkse and Tan (2005) have recently shown that this is incorrect: bids may increase or decrease in n in both private values and common values models. The regression approach might seem more promising in an ascending auction, due to the simplicity of equilibrium bid functions in the button auction model. The multiplicity of equilibria in common values auctions creates one problem. But even ignoring this [e.g., selecting the equilibrium of the button auction given by Milgrom and Weber (1982)] this approach fails due to the fact that the winner's bid is never revealed. For example, in a private values auction the observable bids reveal $(u^{(1:n)}, \ldots, u^{(n-1:n)})$, but $u^{(n:n)}$ is censored. Because the

[93] Note that the second equality need not hold without symmetry. Conditions under which more competition (appropriately defined) implies a more severe winner's curse have not been fully explored.

distribution of the censored valuation $U^{(n:n)}$ varies with n, so does the resulting censoring bias. This makes it impossible to discriminate between private values and common values models based on a regression of bids on n.[94]

In spite of this, and in spite of the lack of identification of many common values models, testing is often possible. Lemma 8.1 makes use of the assumption of exogenous (to the distribution of signals and valuations) variation in the number of bidders. As discussed by Athey and Haile (2002) and Haile, Hong and Shum (2003), this can be reasonable in some applications. Furthermore, it is implied by some models of participation (see Section 6.3.2). However, the assumption of exogenous participation is not always necessary. Initially we will maintain this assumption to simplify the exposition of the basic testing approaches. In Section 8.2 we discuss an approach to testing with endogenous participation.

8.1. Testing in first-price auctions when all bids are observed

In the common values first-price auction, the first-order condition (7.2) requires that $v(x_i, x_i; \mathcal{N}) = \xi(b_i, \mathcal{N})$. Note that both sides of this equation vary with $\mathcal{N}$. However, because $\xi(b_i, \mathcal{N})$ is identified, it is possible to isolate the effect of $\mathcal{N}$ on $v(x_i, x_i; \mathcal{N})$ when $\mathcal{N}$ varies exogenously. Since $v(x_i, x_i; \mathcal{N})$ does not vary with $\mathcal{N}$ in a private values model, it is possible to distinguish the two models, even though $F_{\mathbf{X},\mathbf{U}}(\cdot)$ is not identified. To see how, let $F_{v_i,\mathcal{N}}(\cdot)$ denote the marginal distribution of the random variable $v_i(X_i, X_i; \mathcal{N})$. Lemma 8.1 implies the following result.

COROLLARY 8.1. *Assume exogenous variation in the number of bidders. Then $F_{v_i,\mathcal{N}}(v)$ is invariant to $\mathcal{N}$ in a private values model for all v. In a common values model with symmetric bidders*

$$F_{v_i,\mathcal{N}}(v) < F_{v_i,\mathcal{N}_{+j}}(v) \tag{8.1}$$

for all i, j and all v on the interior of the support of $F_{v_i,\mathcal{N}}(\cdot)$ or $F_{v_i,\mathcal{N}_{+j}}(\cdot)$.

Haile, Hong and Shum (2003) use this result to develop tests of the null hypothesis of private values against the common values alternative.[95] They focus on the case of symmetric bidders, where $F_{v_i,\mathcal{N}}(\cdot)$ can be more simply represented by $F_{v,n}(\cdot)$, and (8.1) can be written

$$F_{v,n}(v) < F_{v,n+1}(v) \quad \forall n, v. \tag{8.2}$$

[94] However, as Bajari and Hortaçsu (2003a) have pointed out, the recurrence relation (8.6) below implies that with this censoring, the average observed bid must increase in n in the dominant strategy equilibrium of a private values button auction. While this is also possible in a common values auction, it provides a testable restriction of the private values hypothesis.

[95] They apply their tests to two types of auctions held by the US Forest Service. Shneyerov (2005) has recently applied one of their tests to data from municipal bond auctions.

Their approach involves two steps. The first is to form estimates $\hat{v}_{it}$ of each $v(x_{it}, x_{it}; n_t)$ using the methods described in Section 3.2. The second step is to compare the empirical distributions

$$\widehat{F}_{\hat{v},n}(v) = \frac{1}{nT_n} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \mathbf{1}\{n_t = n,\ \hat{v}_{it} \leqslant v\}$$

for different values of n.

While tests of equality of distributions (or of the alternative of first-order stochastic dominance) are common in statistics and econometrics, a complication here is the fact that only empirical distributions of the "pseudo-values" $\hat{v}_{it}$ can be compared, not those of the "values" $v(x_{it}, x_{it}; n)$. Hence, the first-stage estimation error (which will be correlated in finite sample for nearby $\hat{v}_{it}$ and $\hat{v}_{jt'}$) must be accounted for. A second complication is the fact that trimming, which must be done separately for each value of n, must be done carefully to avoid creating the appearance of a winner's curse when there is none, or hiding the winner's curse in a common values model.

Haile, Hong and Shum (2003) explore two types of tests.[96] The first is a comparison of trimmed means of each empirical distribution $\widehat{F}_{\hat{v},n}(\cdot)$.[97] For $\tau \in (0, \frac{1}{2})$ let x_τ denote the τth quantile of the marginal distribution $F_X(\cdot)$ and define the quantile-trimmed mean

$$\mu_{n,\tau} = E\big[v(X_i, X_i; n) \mid X_i \in [x_\tau, x_{1-\tau}]\big].$$

Trimming at the same quantiles for all n fixes the set of signals x_i implicitly included in each mean. This is important since the first-order stochastic dominance relation in (8.2) extends to the distributions of $v(X_i, X_i; n)$ over any fixed interval in $[\underline{x}, \overline{x}]$ but need not hold for intervals that vary with n. One can then test the hypotheses

$$\begin{aligned} H_0&: \quad \mu_{\underline{n},\tau} = \cdots = \mu_{\bar{n}}, \\ H_1&: \quad \mu_{\underline{n},\tau} > \cdots > \mu_{\bar{n}}, \end{aligned} \tag{8.3}$$

which are implied by Lemma 8.1.

Let b_τ denote the τth quantile of the observed bids. Since bids are strictly increasing in signals, $\mu_{n,\tau}$ has sample analog

$$\hat{\mu}_{n,\tau} = \frac{1}{nT_n} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{1}\big\{n_t = n,\ b_{it} \in [b_\tau, b_{1-\tau}]\big\}\hat{v}_{it}.$$

Haile, Hong and Shum (2003) show that the vector $(\hat{\mu}_{\underline{n},\tau}, \ldots, \hat{\mu}_{\bar{n},\tau})$ is consistent and has a multivariate normal asymptotic distribution with diagonal covariance matrix Σ, enabling adaptation of a standard multivariate one-sided likelihood-ratio test

[96] Section 5 discusses several other hypotheses to which their tests may be adaptable.

[97] The test generalizes to other finite vectors of functionals – e.g., a vector of quantiles. See Haile, Hong and Shum (2003) for details.

[Bartholomew (1959)]. Monte Carlo evidence suggests that size distortions may be reduced by using the bootstrap to estimate the elements of the covariance matrix Σ.[98]

The second testing approach uses a generalized version of a multi-sample one-sided Kolmogorov–Smirnov test of equal distributions. Given a differentiable strictly decreasing function $\Lambda(\cdot)$, let

$$\Lambda_n(v) = \frac{1}{nT_n} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{1}\{n_t = n\} \Lambda(\hat{v}_{it} - v)$$

and

$$\overline{\delta}_T = \sum_{n=\underline{n}}^{\overline{n}-1} \sup_{v \in [\underline{v}, \overline{v}]} \left[\Lambda_{n+1}(v) - \Lambda_n(v) \right],$$

where the compact interval $[\underline{v}, \overline{v}]$ is bounded away from the endpoints of the support $F_{v,n}(\cdot)$ under the null. If $\Lambda(\cdot)$ is the smoothed step function

$$\Lambda(y) = \frac{\exp(-y/h)}{1 + \exp(-y/h)}$$

with h denoting a bandwidth, $\overline{\delta}_T$ is easily interpreted as an approximation of a more familiar looking one-sided test statistic

$$\delta_T = \sum_{n=\underline{n}}^{\overline{n}-1} \sup_{v \in [\underline{v}, \overline{v}]} \left\{ \widehat{F}_{\hat{v},n+1}(v) - \widehat{F}_{\hat{v},n}(v) \right\},$$

where $\widehat{F}_{\hat{v},n+1}(\cdot)$ and $\widehat{F}_{\hat{v},n}(\cdot)$ denote empirical distribution functions.

Strict monotonicity of $\Lambda(\cdot)$ and the fact that

$$\Lambda_n(v) \to E\left[\Lambda(\hat{v}_{it} - v) \mid n_{it} = n \right]$$

uniformly in $v \in [\underline{v}, \overline{v}]$ imply that $\overline{\delta}_T \to 0$ under the private values null. Under the common values alternative $\overline{\delta}_T \to \delta > 0$. This is the basis for using δ_T as a test statistic. Haile, Hong and Shum (2003) show that for an appropriate normalizing sequence η_T, the generalized Kolmogorov–Smirnov statistic $S_T \equiv \eta_T \overline{\delta}_T$ has a nondegenerate limiting distribution under H_0, enabling use of subsampling for estimation of critical values [e.g., Politis, Romano and Wolf (1999)].

Both types of test are easily adapted to the case in which bidders observe only a signal η of the number of opponents they face before submitting their bids, as long as the econometrician also observes (or can condition on) η. In that case estimation of pseudo-values follows the discussion in Section 6.3.3. One could then compare the distribution of pseudo-values in auctions with higher signals to those with lower signals.

[98] The block bootstrap procedure is identical to that discussed in Section 3.2.1. Haile, Hong and Shum (2003) point out that using the bootstrap to estimate the distribution of the test statistic itself would be difficult, due to the need to resample bids under the null hypothesis on the functions $v(\cdot, \cdot; n)$.

8.2. Testing with endogenous participation

Haile, Hong and Shum (2003) discuss extensions of their testing approaches to cases in which bidder participation is endogenous. If there is a binding reserve price or a cost of preparing a bid, for example, bidders' participation decisions introduce truncation in the set of types submitting bids. They show how their basic approach can still be applied in such cases by comparing estimated distributions of $v(X_i, X_i; n)$, appropriately adjusted for truncation, on regions of common support. We refer readers to their paper for details.

A more difficult case is that in which participation is affected by unobserved factors that also affect valuations. This leads to two quite different threats to the basic testing approach. First, variation in $F_{v,n}(\cdot)$ with n will arise from variation in the unobserved factors, confounding attempts to detect responses to the winner's curse. For example, if auctions of goods that are more valuable in unobserved dimensions also attract greater participation, this could mask the effects of the winner's curse in a common values auction. The second problem is even more fundamental: unobserved heterogeneity threatens the identification of the distributions $F_{v,n}(\cdot)$ that underlies the approach (recall Section 6.1.2).

Haile, Hong and Shum (2003) have proposed an instrumental variables approach for such situations. Consider a simplified version of their approach in which the number of actual bidders in auction t is a function of two scalar factors Z_t and W_t:

$$A_t = \alpha(Z_t, W_t).$$

Here Z_t is an index capturing the effects of factors observable to the econometrician as well as the bidders, while W_t is an index capturing the effects of unobservables.[99] Assume that (i) Z is independent of $(X_1, \ldots, X_{\bar{n}}, U_1, \ldots, U_{\bar{n}})$ and (ii) $\alpha(\cdot,\cdot)$ is weakly increasing in its first argument and strictly increasing in its second.

Assumption (i) is a standard exclusion restriction: Z_t is an instrument affecting participation but not the distribution of valuations and signals. This instrument might be the number of potential bidders or a proxy for it, like the number of firms in the local market. Of course, in principle there need not be any difference between the potential and actual bidders here, based on our definitions. For example, if there is a cost of acquiring a signal but bidders have access to some information about the distribution of valuations before bearing this cost, the number of potential bidders will be correlated with unobservable factors shifting all bidder valuations. Valid instruments in that case might be the number of firms in the market (those who choose whether to invest in a signal), or factors affecting the cost of acquiring a signal.

Assumption (ii) is a monotonicity restriction. Monotonicity in the instrument Z_t implies that changes in Z_t will provide the exogenous variation in the level of competition

[99] For simplicity we assume there are no auction-specific observables other than Z_t, although this is easily relaxed.

that will make it possible to isolate the effects (if any) of the winner's curse. Strict monotonicity in W_t is a key restriction that requires that W_t be discrete (since A_t is). As discussed in Section 6.1.2, this restriction enables identification of the expectations

$$v(x, x; a, z) = E[U_i \mid X_i = x,\ A_t = a,\ Z_t = z]$$

through the first-order condition

$$v(x_{it}, x_{it}; a_t, z_t) = b_{it} + \frac{\Pr(\max_{j\neq i} B_{jt} \leqslant b_{it} \mid B_{it} = b_{it},\ Z_t = z_t,\ A_t = a_t)}{\frac{\partial}{\partial m}\Pr(\max_{j\neq i} B_{jt} \leqslant m \mid B_{it} = b_{it},\ Z_t = z_t,\ A_t = a_t)|_{m=b_{it}}}. \quad (8.4)$$

Estimation of the pseudo-values on the left-hand side of (8.4) proceeds by holding fixed both the value of A and the value of the instrument Z to construct estimators of the right-hand side of (8.4). To test for common values, the pseudo-values $v(x_{it}, x_{it}; a_t, z_t)$ are then pooled across realizations of A_t to compare the cumulative distributions

$$F_{v,z}(v) = \Pr\big(E\big[v(X_{it}, X_{it}; A_t, Z_t)\big] \leqslant v \mid Z_t = z\big)$$

across values of z. While these distributions must be the same for all z under private values, the assumptions above imply that $F_{v,z}(v)$ is increasing in z under the common values alternative. Haile, Hong and Shum (2003) provide additional details and an alternative control function estimation approach allowing for multiple instruments. Their application to US Forest Service timber auctions uses the numbers of sawmills and logging firms in a geographic neighborhood of a sale as instruments for the number of bidders.

8.3. Testing with incomplete bid data

Athey and Haile (2002) show that testing is also possible in ascending auctions (assuming the button auction model) and in first-price auctions in which not all bids are observable.[100] For the symmetric common values model, recall that the challenge arises because the distribution of $v(X^{(n-1:n)}, X^{(n-1:n)}, n)$ varies with n both due to the winner's curse and because the distribution of the order statistic varies with n even without any winner's curse. However, since $X^{(n:n)}$ is unobserved, the distribution of $v(X_i, X_i, n)$ is not identified.

[100] Haile (2001) develops a different testing approach based on detecting bidders' updating of their willingness to pay as an ascending auction proceeds. The insight is that there is no such updating in a private values auction or in a 2-bidder common values auction. Hence one can compare distributions of bidders' willingness to pay (i.e., $\phi(G_B^{(n-1:n)}(\cdot); n-1, n)$) in 2-bidder auctions to that in auctions with larger numbers of bidders, with a difference suggesting common values. A major limitation of this approach is a requirement of independent signals under both the null and alternative hypotheses. While independence is implied by Haile's model of auctions with resale, this will typically be a strong restriction for a common values auction.

Athey and Haile's (2002) approach exploits the fact that for exchangeable random variables $Y_1, \ldots, Y_n$, the marginal distributions $F_Y^{(i:n)}(\cdot)$ of the order statistics must satisfy the recurrence relation [see, e.g., David (1981)]

$$\frac{n-i}{n} F_Y^{(i:n)}(y) + \frac{i}{n} F_Y^{(i+1:n)}(y) = F_Y^{(i:n-1)}(y) \quad \forall y, n, i \leqslant n-1. \tag{8.5}$$

Intuitively, in an *ex ante* sense, moving from a sample of n draws to a sample of $n-1$ draws is equivalent under exchangeability to taking the n draws and then dropping one at random. When one draw, Y_j, is dropped at random from the larger sample, the ith order statistic in the smaller sample will be either the ith order statistic from the larger sample (when Y_j was one of the $n-i$ highest draws), or the $(i+1)$st order statistic (if Y_j was among the i lowest draws). Note that one direct implication of (8.5) is a recurrence relation between means:

$$\frac{n-i}{n} E\big[Y^{(i:n)}\big] + \frac{i}{n} E\big[Y^{(i+1:n)}\big] = E\big[Y^{(i:n-1)}\big] \quad \forall n, i \leqslant n-1. \tag{8.6}$$

Using (8.5) and (8.6), the private values null can be tested against the common values alternative in both first-price and ascending auctions. This is possible even when not all bids are observable (as is always the case in an ascending auction) and despite the fact that the ascending auction has multiple equilibria in the case of common values. The following theorem combines results originally given in Athey and Haile (2002).

THEOREM 8.1. *Assume exogenous variation in the number of bidders. In an ascending auction or first-price sealed-bid auction, the symmetric private values model is testable against the symmetric common values alternative if one observes the bids $B^{(n-2:n)}$ and $B^{(n-1:n)}$ in the ascending auction, or $B^{(n-1:n)}$ and $B^{(n:n)}$ in a first-price auction.*

PROOF. For the first-price auction, we have seen in Theorem 3.3 that the marginal distributions $F_v^{(n-1:n)}(\cdot)$ and $F_v^{(n:n)}(\cdot)$ of $v(X^{(n-1:n)}, X^{(n-1:n)}; n)$ and $v(X^{(n:n)}, X^{(n:n)}; n)$ are identified for all n. In a private values auction, these distributions are $F_U^{(n-1:n)}(\cdot)$ and $F_U^{(n:n)}(\cdot)$ so that (8.5) implies the restriction

$$\frac{1}{n} F_v^{(n-1:n)}(v) + \frac{n-1}{n} F_v^{(n:n)}(v) = F_v^{(n-1:n-1)}(v) \quad \forall v.$$

Under the common values alternative, $v(x, x; n)$ is still a strictly increasing function of x, so that the random variables $v(X_i, X_i; n)$ are still exchangeable. But since $v(X_i, X_i; n)$ strictly decreases in n (Lemma 8.1),

$$\frac{1}{n} F_v^{(n-1:n)}(v) + \frac{n-1}{n} F_v^{(n:n)}(v) > F_v^{(n-1:n-1)}(v)$$

for all v on the interior of the support of $F_{v_i,\mathcal{N}}(\cdot)$ or $F_{v_i,\mathcal{N}_{+j}}(\cdot)$.

For the ascending auction, under the private values null, Equation (8.6) implies

$$\frac{2}{n} E\big[B^{(n-2:n)}\big] + \frac{n-2}{n} E\big[B^{(n-1:n)}\big] = E\big[B^{(n-2:n-1)}\big] \quad \forall n > \underline{n}.$$

Under the common values alternative, Athey and Haile (2002, Theorem 9) show that, regardless of the equilibria selected in the n-bidder and $(n-1)$-bidder auctions, one obtains the relation

$$\frac{2}{n}E\big[B^{(n-2:n)}\big]+\frac{n-2}{n}E\big[B^{(n-1:n)}\big]<E\big[B^{(n-2:n-1)}\big]\quad\forall n>\underline{n}.$$

□

While Theorem 8.1 relies on exchangeability, Athey and Haile (2002) show how this kind of approach can be adapted to asymmetric ascending auctions as well.[101] To see the key idea, observe that if $(Y_1,\ldots,Y_n)$ have an arbitrary joint distribution, one can obtain a sample of exchangeable random variables $(Y_{R_1},Y_{R_2},\ldots,Y_{R_s})$ by taking a random subsample of size $R_s<n$ from the original sample $(Y_1,\ldots,Y_n)$. Hence, even without exchangeability of $(Y_1,\ldots,Y_n)$, a recurrence relation must hold for random subsamples [Balasubramanian and Balakrishnan (1994)]. In a private values auction this implies a recurrence relation between distributions $F_U^{(i:n)}(\cdot)$ in auctions with bidders $\mathcal{N}$ and those from smaller auctions in which the set of bidders is a subset of $\mathcal{N}$.

Formal testing approaches based on these results have not yet been explored. Since the null (alternative) hypothesis can be represented as the hypothesis of equal (stochastically ordered) distributions, it may be possible to adapt the testing approaches of Haile, Hong and Shum (2003), which account for the estimation error arising from the nonparametric estimation of pseudo-values.

8.4. Testing with a binding reserve price

While Haile, Hong and Shum (2003) show that their testing approach can be extended to cases in which there is a binding reserve price, Hendricks, Pinkse and Porter (2003) and Haile, Hong and Shum (2003) have each shown that the presence of a binding reserve price can make possible a different sort of test for the winner's curse in first-price auctions. Both approaches rely on observing the number of potential bidders.

Focusing on the symmetric case, recall that participation is determined by the threshold signal $x^*(n)$ defined by (recall Equation (2.1))

$$x^*(n)=\inf\Big\{x\colon E\Big[U_i\;\Big|\;X_i=x,\ \max_{j\neq i}X_j\leqslant x\Big]\geqslant r\Big\}. \tag{8.7}$$

The equilibrium bid of a bidder with signal $x^*(n)$ is

$$\beta\big(x^*(n);n\big)=v\big(x^*(n),x^*(n);n\big)=E\Big[U_i\;\Big|\;X_i=x^*(n),\ \max_{j\neq i}X_j=x^*(n)\Big].$$

In a private values model, $E[U_i\mid X_i=x,\ \max_{j\neq i}X_j\leqslant x]=E[U_i\mid X_i=x,\ \max_{j\neq i}X_j=x]$, so that

$$\beta\big(x^*(n);n\big)=r. \tag{8.8}$$

[101] They also discuss extension to cases in which only nonadjacent values of n are observed in the data.

As originally noted by Milgrom and Weber (1982), in a common values model the fact that $E[U_i \mid X_i = x,\ \max_{j\neq i} X_j \leqslant x] < E[U_i \mid X_i = x,\ \max_{j\neq i} X_j = x]$ implies

$$\beta\big(x^*(n); n\big) > r. \tag{8.9}$$

Hence, a test for common values can be based on the distinction between (8.8) and (8.9). In particular, if we let $\underline{b} = \inf \operatorname{supp} B_i$,

$$\lim_{b\to\underline{b}} b + \frac{G_{M|B}(b|b; n)}{g_{M|B}(b|b; n)}$$

should equal r under the private values hypothesis but should be strictly greater than r with common values. While this idea was first mentioned by Hendricks, Pinkse and Porter (2003), a formal test based on this idea has not yet been developed.

A second possibility, suggested by Haile, Hong and Shum (2003), is to examine variation in the probability $F_X(x^*(n))$ that the reserve price excludes a potential bidder. It is easy to verify (following the proof of Lemma 8.1) that $x^*(n)$ is invariant to n in a private values model but strictly increasing in n in a common values model. By exchangeability,

$$F_X\big(x^*(n)\big) = F_{\mathbf{X}}\big(x^*(n), \infty, \ldots, \infty\big) = \sum_{k=1}^{n} \frac{k}{n} \Pr(A = n - k \mid N = n).$$

So if both the number of potential bidders, N, and the number of actual bidders, A, are observed, $F_X(x^*(n))$ is identified for all n, and one can test whether this is constant or decreasing in n.

9. Dynamics

Until very recently, virtually all structural empirical work on auctions has considered static models, treating each auction in the data as an independent game. There are several reasons this may not be the case. First, even in a stationary environment, dynamic considerations arise if firms engage in collusion.[102] We do not consider collusion in this chapter.[103] Second, bidders' valuation distributions may change over time in a way that

[102] Many models of collusion at auctions are static [e.g. McAfee and McMillan (1992)]. Recently, the theory of tacit collusion in repeated auctions has grown rapidly [Aoyagi (2003), Athey and Bagwell (2001, 2004a), Athey, Bagwell and Sanchirico (2004) and Skryzpacz and Hopenhayn (2004)]. Athey, Bagwell and Sanchirico (2004) show that when only the winning bid, but not the bidder's identity, is revealed by the auctioneer, optimal collusion entails bidding at the reserve price with each bidder having an equal chance of winning, while Athey and Bagwell (2001, 2004a) show that when the bidder's identity is revealed as well, bidders engage in sophisticated rotation schemes so that a bidder's probability of winning is less correlated over time than the bidder's valuation.

[103] For empirical studies of collusion (which typically do not explicitly consider dynamics), see Porter and Zona (1993, 1999), Bajari and Ye (2003), Pesendorfer (2000), Baldwin, Marshall and Richard (1997), and Athey, Levin and Seira (2004). See Bajari and Summers (2002) for a survey.

is exogenous, but is private information to each bidder.[104] This can create dynamic links in bidder strategies. In particular, a bidder's behavior in an auction will affect opponents' beliefs about his valuation distribution in future auctions, changing the equilibrium of the auction game in each period.[105] To our knowledge, there has been no empirical investigation focusing on the dynamics of such models.

Finally, the underlying distribution of valuations might change as a function of auction outcomes, potentially in ways that are observable (or can be directly inferred) by the other bidders. For example, there may be learning-by-doing, so that a firm that wins an auction today might have stochastically lower costs (higher valuations) in the future. Alternatively, firms may have capacity constraints (or more general forms of diseconomies of scale). In that case, a firm that wins an auction today might draw a valuation from a less favorable distribution in the future. In either case, the resulting dynamic considerations for bidders will change the equilibrium at each point in time.

To explore this type of environment, consider a model based on that of Jofre-Bonet and Pesendorfer (2000, 2003).[106] Time is discrete, and firms compete over an infinite horizon. In each period t, an item is sold by first-price auction to one of n bidders. For simplicity, assume that there is no reserve price and that all objects to be auctioned have the same observable characteristics [see Jofre-Bonet and Pesendorfer (2003) for extensions]. The distribution of bidder valuations depends on the bidders' capacities (more generally, it could depend on other covariates as well). Conditional on capacities, bidder valuations are independent across bidders and over time. Letting $c_{i,t}$ be bidder i's publicly observable capacity in period t, the conditional distribution of bidder i's valuation in period t is denoted $F_U(\cdot|c_{i,t})$, where for simplicity we let this function be the same for all bidders.

The econometrician and the bidders both know the (deterministic) transition function for bidder capacities. In particular, if k is the identity of the winning bidder in period t and $\mathbf{c}_t$ is the vector of bidder capacities in period t, then[107]

$$c_{i,t+1} = \omega_i(\mathbf{c}_t, k).$$

The solution concept is Markov-perfect equilibrium. Thus, collusion is ruled out, and dynamic considerations in bidder strategies arise only because bidders anticipate that the identity of today's winner will affect future capacities, which in turn will affect

[104] If the distribution of valuations changes over time in a way that is observed by all bidders, then either the econometrician can observe (and condition on) the factors affecting distribution, or the problem of unobserved heterogeneity discussed in Section 6.1.2 arises. The literature has not explored intertemporal correlation in unobserved auction heterogeneity.

[105] See e.g., Bikhchandani (1988), Bikhchandani and Huang (1989), Haile (1999, 2003), Katzman and Rhodes-Kropf (2002), Das Varma (2003), Goeree (2003), and Athey and Bagwell (2004b).

[106] They analyze a procurement auction. We recast the problem as one in which the bidders are buyers. They also consider two types of bidders, "regular bidders" who bid often, and "fringe bidders" who bid rarely and do not consider the future. We focus on regular bidders to simplify the analysis.

[107] Jofre-Bonet and Pesendorfer (2003) allow a slightly richer specification in which transitions reflect information about the size and duration of projects that have been won in the past.

outcomes in future auctions. Since all asymmetries are captured through capacities, Jofre-Bonet and Pesendorfer (2003) focus on exchangeable strategies. In particular, each bidder's bid in a given period depends on the bidder's own valuation and the vector of capacities, so that strategies can be written $\beta_i(u_{i,t}, \mathbf{c}_t)$.[108]

In this environment, Jofre-Bonet and Pesendorfer (2003) combine the insights of Hotz and Miller (1993) (who studied dynamic discrete choice problems for individuals) with the approach of Guerre, Perrigne and Vuong (2000) in an insightful way, providing very general conditions for identification when the discount factor is known.

The first step in the analysis is to use dynamic programming to represent bidder payoffs.[109] Suppressing $\mathcal{N}$ in the notation, let $G_{M_i}(\cdot|\mathbf{c})$ be the equilibrium distribution of the maximum opponent bid for bidder i when the vector of bidder capacities is $\mathbf{c}$. Let δ denote the discount factor and let $\omega(\mathbf{c}, k) = (\omega_1(\mathbf{c}, k), \ldots, \omega_n(\mathbf{c}, k))$. Holding opponents' strategies fixed, the interim expected discounted sum of future profits for bidder i is given by

$$W_i(u_i, \mathbf{c}) = \max_{b_i} \Bigg\{ (u_i - b_i) G_{M_i}(b_i|\mathbf{c}) + \delta \sum_{j=1}^{n} \Pr(j \text{ wins } |b_i, \mathbf{c}) \int_{u_i'} W_i\big(u_i', \omega(\mathbf{c}, j)\big) f_{U_i}\big(u_i' \mid \omega_i(\mathbf{c}, j)\big)\, \mathrm{d}u_i' \Bigg\},$$

where the second term sums over the possible identities of the winner to form an expectation of the continuation value to player i, given current capacities. One can then define the *ex ante* value function

$$V_i(\mathbf{c}) = \int W_i(u_i, \mathbf{c}) f_{U_i}(u_i|\mathbf{c})\, \mathrm{d}u_i,$$

which can be rewritten as

$$V_i(\mathbf{c}) = \int \Bigg\{ \max_{b_i} \Bigg\{ (u_i - b_i) G_{M_i}(b_i|\mathbf{c}) + \delta V_i\big(\omega(\mathbf{c}, i)\big) + \delta \sum_{j \neq i} \Pr(j \text{ wins } |b_i, \mathbf{c}) \big[V_i\big(\omega(\mathbf{c}, j)\big) - V_i\big(\omega(\mathbf{c}, i)\big) \big] \Bigg\} \Bigg\} f_{U_i}(u_i|\mathbf{c})\, \mathrm{d}u_i.$$

Note that in equilibrium, the probability that bidder i wins with bid b_i is given by

$$\Pr\Big(b_i \geqslant \max_{j \neq i} \beta_j(U_j, \mathbf{c}) \;\Big|\; \mathbf{c} \Big) = G_{M_i}(b_i|\mathbf{c}) = \prod_{j \neq i} G_{B_j}(b_i|\mathbf{c}), \tag{9.1}$$

[108] Jofre-Bonet and Pesendorfer (2000, 2003) establish existence of an equilibrium within the parametric framework they use for estimation, and also sketch an approach for showing existence in general.

[109] Ackerberg et al. (Chapter 63 in this volume) discuss estimation of dynamic strategic models more generally, which relies on very similar ideas.

where $G_{B_j}(\cdot|\mathbf{c})$ is the cumulative distribution of B_j conditional on capacity vector $\mathbf{c}$. The probability that bidder $j \neq i$ wins when bidder i bids b_i is

$$\int_{b_i}^{\bar{b}_j(\mathbf{c})} \Bigg(\prod_{k \neq i,j} G_{B_k}(b_j|\mathbf{c}) \Bigg) g_{B_j}(b_j|\mathbf{c})\, \mathrm{d}b_j,$$

where $\bar{b}_j(\mathbf{c}) = \sup \operatorname{supp} G_{B_j}(\cdot|\mathbf{c})$. Finally, using (9.1) note that

$$\frac{G_{M_i}(b_i|\mathbf{c})}{g_{M_i}(b_i|\mathbf{c})} = \frac{1}{\sum_{j \neq i} \frac{g_{B_j}(b_i|\mathbf{c})}{G_{B_j}(b_i|\mathbf{c})}}.$$

The next step is to solve for the *ex ante* value functions in terms of observables. This requires a significant generalization of the two-step indirect approach proposed by Guerre, Perrigne and Vuong (2000). Consider bidder i's optimization problem in a given auction:

$$\max_{b_i} \Bigg\{ (u_i - b_i) G_{M_i}(b_i|\mathbf{c}) + \delta V_i\big(\omega(\mathbf{c}, i)\big) + \delta \sum_{j \neq i} \Bigg(\int_{b_i}^{\bar{b}_j(\mathbf{c})} \prod_{k \neq i,j} G_{B_k}(b_j|\mathbf{c}) g_{B_j}(b_j|\mathbf{c})\, \mathrm{d}b_j \Bigg) \big[V_i\big(\omega(\mathbf{c}, j)\big) - V_i\big(\omega(\mathbf{c}, i)\big) \big] \Bigg\}.$$

The first-order condition is

$$u_i = b_i + \frac{G_{M_i}(b_i|\mathbf{c})}{g_{M_i}(b_i|\mathbf{c})} + \delta \sum_{j \neq i} \frac{G_{M_i}(b_i|\mathbf{c})}{g_{M_i}(b_i|\mathbf{c})} \frac{g_{B_j}(b_i|\mathbf{c})}{G_{B_j}(b_i|\mathbf{c})} \big(V_i\big(\omega(\mathbf{c}, j)\big) - V_i\big(\omega(\mathbf{c}, i)\big) \big). \tag{9.2}$$

After substituting this into the *ex ante* value function, a change of variables yields

$$\begin{aligned} V_i(\mathbf{c}) = {} & \int_{\underline{b}_i(\mathbf{c})}^{\bar{b}_i(\mathbf{c})} \frac{G_{M_i}(b_i|\mathbf{c})}{g_{M_i}(b_i|\mathbf{c})} G_{M_i}(b_i|\mathbf{c})\, \mathrm{d}G_{B_i}(b_i|\mathbf{c}) \\ & + \delta \sum_{j \neq i} V_i\big(\omega(\mathbf{c}, j)\big) \Bigg\{ \int_{b_i}^{\bar{b}_i(\mathbf{c})} \prod_{k \neq i,j} G_{B_k}(b_j|\mathbf{c}) g_{B_j}(b_j|\mathbf{c})\, \mathrm{d}b_j \\ & + \int_{\underline{b}_i(\mathbf{c})}^{\bar{b}_i(\mathbf{c})} \frac{G_{M_i}(b_i|\mathbf{c})}{g_{M_i}(b_i|\mathbf{c})} \frac{g_{B_j}(b_i|\mathbf{c})}{G_{B_j}(b_i|\mathbf{c})} G_{M_i}(b_i|\mathbf{c})\, \mathrm{d}G_{B_i}(b_i|\mathbf{c}) \Bigg\}. \end{aligned}$$

For any capacity vector $\mathbf{c}$, this expresses each $V_i(\mathbf{c})$ as a linear function of $V_i(\cdot)$ evaluated at other capacity vectors. The coefficients of this linear relation depend only on the observable bid distributions. Thus, it is possible to solve for the *ex ante* value functions in terms of the observable bid distributions.

Once the *ex ante* value functions have been computed, identification of the distributions $F_U(\cdot|\mathbf{c})$ (assuming the discount factor δ is known) follows from the first-order

condition (9.2). In particular, we can use the observed bid distributions and the *ex ante* value functions to compute the right-hand side of (9.2). Then, if the discount factor δ is known (for example, from other empirical studies), (9.2) implies that $F_{U_i}(\cdot|\mathbf{c})$ is identified from the observed bids and capacities.

In addition to demonstrating the nonparametric identification of their model, Jofre-Bonet and Pesendorfer (2003) propose a parametric estimation approach, motivated in part by a desire to include covariates in a parsimonious manner. To solve for the value functions, they follow Judd (1998) and discretize the set of possible capacities. Then, calculating the value functions entails solving a system of linear equations. They further simplify the estimation by using a quadratic approximation of the value function. They apply their approach to California highway construction contracts. Using their estimates, they are able to assess the importance of private information, capacity constraints, and the inefficiencies that arise due to the asymmetries induced by capacity differences among bidders under the assumption of forward-looking equilibrium behavior. Note that it is impossible to *test* whether bidders are forward looking in this environment, since the discount factor δ is not identified.

10. Multi-unit and multi-object auctions

10.1. Auctions of perfect substitutes

While most of the empirical literature on auctions focuses on the case of single-unit auctions, auctions of multiple units of identical goods ("multi-unit auctions") have recently begun to gain significant attention. One motivation is their importance in the public sector. For example, multi-unit auctions have recently been implemented in restructured electricity markets to assign electric power generation to different plants [see, e.g., Wolfram (1998), Borenstein, Bushnell and Wolak (2002), or Wolak (2003)]. Optimal design of such markets is complex: the usual goals of efficiency and surplus extraction in single-unit auctions are complicated by (among other issues) nonlinearities in cost functions, incentives to exercise market power by withholding marginal production capacity, and the need for firms to recover substantial fixed costs [Wolak (2003)]. Empirical analysis of these markets can provide valuable information about the underlying cost structure, market power opportunities, and profitability. Another important policy question that has been the subject of discussion among economists at least since Friedman (1960) is how governments should auction treasury securities to maximize revenues. This question is potentially relevant to the design of markets for other types of securities as well.

In treasury auctions, a large number of identical securities is sold in a mechanism in which each bidder submits an entire "demand function," i.e., each bidder i offers a (downward sloping) schedule of price-quantity combinations (b_{ij}, q_j) specifying the

price he is willing to pay for his q_jth unit.[110] Two auction mechanisms are commonly used: *discriminatory* and *uniform-price*. A discriminatory auction is the most common in practice (although recently the US adopted uniform-price auctions after conducting an experiment to evaluate alternative formats). In this mechanism, each bidder who offers more than the market clearing bid for a unit receives that unit at the price he offered. As the name suggests, this results in different prices for different units of the same security – even a given bidder will pay different prices for each unit he wins. In contrast, in a uniform-price auction, the market clearing price (lowest accepted bid) is paid on all units sold. In addition to US treasury bill auctions, electricity auctions are often uniform price, and some firms have used uniform price auctions in initial public offerings.[111]

Of course, bidders will bid differently depending on whether a discriminatory or uniform-price auction is used. The revenue ranking of the two mechanisms is theoretically ambiguous [Ausubel and Cramton (2002)] and can only be determined with knowledge of the true distribution of bidder valuations. To our knowledge, the literature has not yet presented a comprehensive analysis of identification and estimation in uniform-price auctions, although Wolak (2003) provides some initial results.

Before proceeding, we pause to highlight the fact that the theory of multi-unit auctions is much less well developed than the theory of first-price auctions and ascending auctions. Although existence of equilibrium in mixed strategies can be guaranteed quite generally [Jackson et al. (2002), Jackson and Swinkels (2005)], existence of pure strategy Nash equilibria in monotone strategies has been established for only a limited class of models, such as private or common value models where bidder signals are independent [McAdams (2004a)]. In addition, examples have shown that there can be multiple equilibria [e.g. Back and Zender (1993)]. Thus, most existing econometric approaches to these auctions require assumptions on endogenous variables to guarantee that the requisite regularity properties are satisfied, although in practice some of the conditions can be verified empirically.

Hortaçsu (2002) has empirically analyzed the question of which auction mechanism raises higher revenue, and has shown that the relevant primitives can be identified nonparametrically in a private values model of the discriminatory auction.[112] His empirical model is based on the theoretical framework of Wilson (1979).[113] Building on the insight of Guerre, Perrigne and Vuong (2000), he points out that equilibrium bidding

[110] Note that this is a bidder's *strategic* expression of quantities he demands at each price. This need not correspond to the usual notion of a price-taking buyer's demand function.

[111] In the finance literature, these are often referred to as "Dutch auctions," conflicting with economists' use of this term for descending price single-unit auctions.

[112] Parametric structural models have been studied recently by Février, Préget and Visser (2002) and by Armantier and Sbaï (2003), both of which consider common values models. Common values models may be appropriate for many securities auctions, although this is ultimately an empirical question – one for which testing approaches have not been developed. Hortaçsu (2002) discusses institutional details that motivate the assumption of private values in the case of the Turkish treasury bill auctions he studies.

[113] The analysis in Wilson's model relies on an assumption that bidders can bid continuous demand functions. In most applications, bids are restricted (by rule or in practice) to step functions – i.e., finite sets of discrete

strategies can be characterized as best responses by each bidder to the distribution of opposing bids he faces. In this multi-unit setting, the distribution of opposing bids cannot be described by the distribution of the maximum opposing bid (as in a single-unit auction); rather, it is the stochastic *residual supply curve* that characterizes the equilibrium probabilities with which various quantities could be obtained at each possible price.

For a discriminatory auction, suppose that the total quantity of securities to be offered is Q. Let $q_i(\cdot)$ denote the demand function offered by bidder i; i.e., $q_i(p)$ is the largest quantity for which bidder i is offering a price of p or more for his final unit. For a given set of demand functions $q_1(\cdot), \ldots, q_n(\cdot)$, the market clearing price p^c then equates supply and demand:

$$Q = \sum_i q_i\big(p^c\big).$$

This market clearing price can be reinterpreted as the price at which i's demand function and his residual supply curve intersect:

$$Q_{R_i}(b) = Q - \sum_{j \neq i} q_j(b).$$

Let $v_i(y; x_i)$ denote bidder i's marginal valuation for a yth unit of the good, given the realization of his signal x_i. Each bidder i's equilibrium strategy specifies, for each possible realization of x_i, a demand function $q_i(b) = \varphi_i(b; x_i)$ expressing the quantity demanded at each price b. Let

$$G_i(b, y) = \Pr\left(y \leqslant Q - \sum_{j \neq i} \varphi_j(b; X_j)\right) \tag{10.1}$$

so that $G_i(b, y)$ is the probability that, given equilibrium bidding by i's opponents, the market clearing price falls below b if i himself demands quantity y at price b.

For each $X_i = x_i$, bidder i's optimal strategy $\varphi_i(\cdot; x_i)$ then solves the problem

$$\max_{q_i(\cdot)} \int_0^\infty \left(\int_0^{q_i(p^c)} \big(v_i(y, x_i) - q_i^{-1}(y)\big)\, \mathrm{d}y \right) \frac{\partial G_i(p^c, q_i(p^c))}{\partial p^c}\, \mathrm{d}p^c.$$

One can show that the optimal bidding strategy can be characterized by the necessary condition

$$v_i\big(\varphi(b; x_i), x_i\big) = b + \frac{G_i(b, \varphi(b; x_i))}{\frac{\partial}{\partial b} G_i(b, \varphi(b; x_i))}.$$

While this is an Euler–Lagrange condition for a functional optimization problem, this equation closely resembles the first-order condition (2.4) used by Guerre, Perrigne and

price-quantity pairs. Recently, Wolak (2004), McAdams (2005), and Kastl (2005) have explored empirical models explicitly accounting for this discreteness.

Vuong (2000) to show identification of the single-unit discriminatory (i.e., first-price sealed-bid) auction with private values. Its role in the identification argument is similar. Because the demand functions $q_j(b) = \varphi_i(b; x_j)$ are directly observed, $G_i(b, y)$ is identified from Equation (10.1). Then, for any quantity y demanded at price b by bidder i, we have

$$v_i(y, x_i) = b + \frac{G_i(b, y)}{\frac{\partial}{\partial b} G_i(b, y)},$$

which uniquely determines the realizations of bidder i's marginal valuations at each quantity y. This implies identification of the distributions of each $v_i(y, X_i)$, which are the primitives needed for policy simulations.[114] In particular, if for each quantity y we let B_i^y be a random variable equal in distribution to $\varphi_i^{-1}(y; X_i)$, $v_i(y, X_i)$ must be equal in distribution to

$$B_i^y + \frac{G_i(B_i^y, y)}{\frac{\partial}{\partial b} G_i(b, y)|_{b=B_i^y}}.$$

Hortaçsu (2002) explores several estimation approaches, both parametric and nonparametric. He also finds a clever way to place an upper bound on the revenue that would be obtained under the uniform price auction, while avoiding the difficult problem of solving for the equilibrium given the estimated distribution of valuations: since a bidder will never bid more than her marginal valuation for each unit, the revenue that would be obtained if bidders simply bid their marginal valuations for each unit in a uniform auction provides an upper bound on the equilibrium revenue.

10.2. Auctions of imperfect substitutes and complements

One prominent area in which economists' understanding of auctions has been used to guide policy over the last decade is in the design of institutions to allocate spectrum rights [see, e.g., McAfee and McMillan (1996)]. Questions regarding the optimal design of spectrum auctions led to much new theoretical work considering the complications to equilibrium strategies arising in multi-object auctions, where the heterogeneous goods auctioned at the same time may be imperfect substitutes, complements, or combinations of these. Similar issues arise in a number of procurement applications, where complementarities may exist between contracts, and some bundles of contracts may be substitutes for others. Cantillon and Pesendorfer (2003) study one such application: auctions for bus services in London, where it may be cheaper to operate one route if a nearby route is also served. Here, we describe their model and identification results. For

[114] Note that signals play a purely informational role here. Hence, their distribution can be normalized (and assumed symmetric) without loss of generality. Put differently, only the distribution of marginal valuations, not that of the underlying signals, is of economic relevance.

consistency with the rest of the chapter, we treat the auction as one in which the bidders are buyers rather than sellers of services.

Let S be the set of goods offered for sale, with $|S| = m$. Let $U_{i,s}$ be bidder i's valuation for the bundle $s \subseteq S$, with $\mathbf{U}_i \in \mathbb{R}^{2^m-1}$ denoting the vector of his valuations for all possible bundles $s \subseteq S$. Bidders' preferences over combinations of goods may exhibit sub- and/or super-additivity. Let $F_{\mathbf{U}_i}(\cdot)$ be the joint distribution of $\mathbf{U}_i$, while $F_{U_{i,s}}(\cdot)$ denotes the marginal distribution of $U_{i,s}$. Let $B_{i,s}$ denote bidder i's bid on bundle s, and let $\mathbf{B}_i$ be the vector of bids placed by bidder i. We let $\mathbf{B}_{i,-s}$ denote the vector of bids placed by bidder i on all bundles other than s.

For simplicity, we focus on a fixed set of n symmetric bidders. Bidders participate in a sealed-bid discriminatory auction with *combination bidding*: each bidder submits bids on all bundles, and the auctioneer chooses the allocation of all objects that maximizes total revenue, charging each bidder the price he offered for each bundle he is allocated.

Combination bidding enables bids to express complementarities and substitutabilities between objects and/or bundles. Further, we might expect combination bidding to aid efficiency and to encourage less cautious bidding by bidders who desire certain combinations of goods. However, combination bidding also introduces a strategic incentive absent in auctions of homogeneous goods. This arises from the fact that a bidder's bid on one bundle competes with his own bids on other bundles. If a bidder raises his bid for bundle s, for example, that will make him more likely to win s, but it may reduce his chances of winning a different bundle t. This is because an increase in $b_{i,s}$ may make it profitable for the seller to allocate bundle s to i instead of bundle t, allocating t to some other bidder instead.

A bidder's problem here turns out to be very closely related to the problem of multiproduct pricing, where it is known that a firm may find it profitable to bundle goods for which demands are independent. Analogously, here a bidder may find it profitable to place bids on bundles (i.e., to make "combination bids" or "bundle bids") even if the goods in the bundle are independent in the sense that $U_{i,s\cup t} = U_{i,s} + U_{i,t}$ when $s \cap t = \emptyset$. This is because the combination bid on the bundle $s \cup t$ can enable bidder i to win bundle s even when bidder i's opponents place a high bid for bundle s, unless they also place a high bid for bundle t. Thus, the combination bid allows bidder i to "leverage" a high valuation for bundle s into a lower price paid for bundle t, or vice versa [cf. Whinston (1989)]. Note that this leads bidder i to bid less aggressively on the individual bundles s and t, in order to avoid competing with her combination bid.

Following intuition from the literature on bundling [see McAfee, McMillan and Whinston (1989) or Armstrong and Rochet (1999)], as long as the correlation among opponent bids for s and t is not too high, making a combination bid is profitable for the bidder. Cantillon and Pesendorfer (2003) describe a plausible environment in which allowing combination bids will reduce both expected revenue and efficiency if goods are independent. This provides one motivation for determining whether bidders view the goods as independent, substitutes, or complements.

For the purposes of this section, we will make the following nonprimitive assumptions (Cantillon and Pesendorfer use slightly weaker assumptions)[115]: a pure strategy Nash equilibrium exists, the joint distribution of equilibrium bid vectors $(\mathbf{B}_1, \ldots, \mathbf{B}_n)$ is differentiable almost everywhere in the support of equilibrium bids, and there is zero probability that bidder i uses a bid in equilibrium at which the joint distribution of opponent bids fails to be differentiable.

Given the equilibrium distribution of bid vectors for bidder i's opponents $j \neq i$, let $G^s(\mathbf{b}_i)$ denote the probability that bidder i wins the objects in bundle s when bidder i chooses the bid vector $\mathbf{b}_i$. Note that $G^s(\cdot)$ generally is not a cumulative distribution function and need not even be increasing, since increasing $b_{i,t}$ for a bundle t such that $s \cap t \neq \emptyset$ might lead to a lower probability that i wins all objects in bundle s. When there are no reserve prices, bidder i solves the problem

$$\max_{\mathbf{b}_i} \sum_{s \subseteq S} (u_{i,s} - b_{i,s}) G^s(\mathbf{b}_i).$$

If $\mathbf{b}_i$ is the equilibrium bid for bidder i when his type is $\mathbf{u}_i$, then as long as the objective function is differentiable at $\mathbf{b}_i$, the following system of first-order conditions must be satisfied:

$$-G^s(\mathbf{b}_i) + \sum_{t \subseteq S} (u_{i,t} - b_{i,t}) \frac{\partial}{\partial b_{i,s}} G^t(\mathbf{b}_i) = 0 \quad \text{for all } s \subseteq S. \tag{10.2}$$

Let $\mathbf{G}(\mathbf{b}_i)$ denote the $(2^m - 1) \times 1$ vector with components $G^s(\mathbf{b}_i)$, and let $\nabla \mathbf{G}(\mathbf{b}_i)$ be the $(2^m - 1) \times (2^m - 1)$ matrix with (s, t) element $\frac{\partial}{\partial b_{i,s}} G^t(\mathbf{b}_i)$. Then we can rewrite the system of first-order conditions in matrix notation as

$$\nabla \mathbf{G}(\mathbf{b}_i)[\mathbf{u}_i - \mathbf{b}_i] = \mathbf{G}(\mathbf{b}_i).$$

This is a system of linear equations in the vector of valuations $\mathbf{u}_i$. If $\nabla \mathbf{G}(\mathbf{b}_i)$ is invertible, we can rewrite the first-order conditions in a form analogous to the single-unit auction case (2.4):

$$\mathbf{u}_i = \mathbf{b}_i + \left[\nabla \mathbf{G}(\mathbf{b}_i)\right]^{-1} \mathbf{G}(\mathbf{b}_i). \tag{10.3}$$

Invertibility of $\nabla \mathbf{G}(\mathbf{b}_i)$ would then imply that the distribution of (multidimensional) valuations were nonparametrically identified, following the logic developed above for the single-object first-price auction.

[115] They argue that a mixed strategy equilibrium exists, but to our knowledge it is not known what additional assumptions would be required to guarantee that a pure strategy equilibrium exists. Although it might seem that a mixed strategy equilibrium should be inconsistent with identification, that is not necessarily true. In a mixed strategy equilibrium, for at least some valuations, a bidder uses more than one bid vector: the mapping from valuations to bids is one-to-many. Identification of the primitive valuation functions will require that for each bid vector, there is a unique valuation that uses that bid vector; i.e, that the mapping from bids to valuations is many-to-one.

One unresolved question is whether there are useful sufficient conditions on the distribution of bids (or on $\mathbf{G}(\cdot)$) that ensure that observed bidding is consistent with equilibrium behavior (see Section 5.1). First-order conditions are, of course, necessary but not sufficient for equilibrium. In the case of a single-unit first-price auction, Theorem 5.2 ensures that the first-order conditions together with monotonicity of the (inverse) bid function are necessary and sufficient for optimality of each bidder's best response. Thus far there is no analogous result for the multi-object auction considered here. Hence, for an observed bid vector $\mathbf{b}_i$ it is possible that there is a unique $\mathbf{u}_i$ satisfying (10.3), yet for that $\mathbf{u}_i$, $\mathbf{b}_i$ is not a best response to the distribution of i's opponents' bids. However, it should be possible to rule this out in a given application: since the bidder's objective function can be calculated from observables for each vector of valuations, for each observed $\mathbf{b}_i$ and corresponding $\mathbf{u}_i$ satisfying (10.3) it is possible to compute the globally optimal bid vector for $\mathbf{u}_i$ and confirm that it is equal to $\mathbf{b}_i$, thereby verifying that the inverse bid functions implied by (10.3) are mutual best responses.

A second difficulty with using (10.3) arises from the fact $\nabla\mathbf{G}(\cdot)$ will not in general be invertible, since bidders need not make bids on all bundles – not even on all those for which they have positive valuations. Making no bid on a given bundle (or, equivalently, making a bid for this bundle that is sure to lose) can be optimal for a bidder since this ensures that she does not compete with her own bids on other bundles. Given a bid vector $\mathbf{b}_{i,-s}$, Cantillon and Pesendorfer (2003) call a bid $b_{i,s}$ *irrelevant* if

$$b_{i,s} < \inf\bigl\{\tilde{b}_{i,s}\colon\ G^s(\tilde{b}_{i,s}, \mathbf{b}_{i,-s}) > 0\bigr\}.$$

Irrelevant bids are bids that could never win. The problem for identification is that if a bidder places an irrelevant bid on bundle s, $\frac{\partial}{\partial b_{i,t}} G^s(b_i) = 0$ and $\frac{\partial}{\partial b_{i,s}} G^t(b_i) = 0$ for all $t \subseteq S$, implying that $\nabla\mathbf{G}(\mathbf{b}_i)$ is not invertible. Indeed, Cantillon and Pesendorfer (2003) establish that $\nabla\mathbf{G}(\mathbf{b}_i)$ is invertible if and only if there are no irrelevant bids. In their application, bidders appear to make many irrelevant bids.[116]

Although irrelevant bids preclude point identification, there is still information in such bids. First observe that if $\mathbf{b}_i$ includes an irrelevant bid for bundle t, it is still possible to identify the valuations associated with the bids for other bundles. To see this, note that if for valuation vector $\mathbf{u}_i$ it is optimal to place relevant bids for all bundles in $K \subset 2^S$ and irrelevant bids on other bundles, one obtains the same solution if one treats the bidder's optimization problem as a constrained problem, with the bidder *required* to place irrelevant bids on all bundles $2^S \backslash K$. Formally, let $\mathbf{b}_i^K$ be the subvector of bids on the elements of K, and let $G_K^s(\mathbf{b}_i^K)$ denote the probability that bidder i wins bundle s when he places irrelevant bids on bundles $t \in 2^S \backslash K$ and bids $\mathbf{b}_i^K$ on bundles in K. Finally, let $\mathbf{G}_K(\cdot)$ denote a vector with elements given by $G_K^s(\cdot)$ for $s \in K$. Then the optimal bid for type $\mathbf{u}_i$ of bidder i in the original game is also the solution to

$$\max_{\mathbf{b}_i^K} \sum_{s \subseteq K} (u_{i,s} - b_{i,s}) G_K^s\bigl(\mathbf{b}_i^K\bigr).$$

[116] Irrelevant bids are identified by replacing $\mathbf{G}(\cdot)$ with the empirical analog and directly checking whether each bid has a positive probability of winning.

The solution to this problem will involve no irrelevant bids, so $\nabla \mathbf{G}_K(\mathbf{b}_i^K)$ will be invertible. Hence, the valuations $\mathbf{u}_i^K$ that in equilibrium correspond to bids $\mathbf{b}_i^K$ will be identified.

There is also information in bids about valuations for bundles for which irrelevant bids have been placed. Given a bid vector $\mathbf{b}_{i,-s}$, define the "effective bid"

$$b_{i,s}^{\text{eff}} = \inf\bigl\{b_s\colon\ G^s(b_s, \mathbf{b}_{i,-s}) > 0\bigr\}.$$

Given continuity of payoffs and the opponent bid distribution, bidder i will always be indifferent between bidding $(b_{i,s}, \mathbf{b}_{i,-s})$, where $b_{i,s}$ is irrelevant, and $(b_{i,s}^{\text{eff}}, \mathbf{b}_{i,-s})$. This implies that increasing $b_{i,s}$ is unprofitable at $b_{i,s} = b_{i,s}^{\text{eff}}$ when $b_{i,s}$ is irrelevant, i.e.,

$$\begin{aligned}&\frac{\partial}{\partial b_{i,s}} G^s(b_{i,s}, \mathbf{b}_{i,-s})\bigg|_{b_{i,s}=b_{i,s}^{\text{eff}}} \bigl(u_{i,s} - b_{i,s}^{\text{eff}}\bigr)\\ &\quad + \sum_{t\subseteq S, t\neq s} (u_{i,t} - b_{i,t}) \frac{\partial}{\partial b_{i,s}} G^t(b_{i,s}, \mathbf{b}_{i,-s})\bigg|_{b_{i,s}=b_{i,s}^{\text{eff}}} \leqslant 0 \quad \text{for all } s \subseteq S, \end{aligned} \tag{10.4}$$

where all derivatives are taken from the right. Since $\frac{\partial}{\partial b_{i,s}} G^t(b_{i,s}^{\text{eff}}, \mathbf{b}_{i,-s}) = 0$ (again taking the derivative from the right) for all $t \neq s$ such that $b_{i,t}$ is irrelevant, and since we have just argued that $u_{i,t}$ is identified for all t such that $b_{i,t}$ is relevant, the only remaining unknown in (10.4) is $u_{i,s}$. Thus, (10.4) places an upper bound on the bidder's valuation for bundle s. In particular, the true $u_{i,s}$ must be less than the value of $u_{i,s}$ that makes (10.4) hold with equality. This can be used to provide a lower bound on the cumulative distribution of $U_{i,s}$. More generally, a lower bound on the distribution of $\mathbf{U}_i$ is identified using (10.4).

In Cantillon and Pesendorfer's application, two additional constraints are imposed on bids. First, there are reserve prices, denoted r_s; bids below the reserve price win with probability zero. Second, the auction rules specify that

$$b_{i,s\cup t} \geqslant b_{i,s} + b_{i,t} \quad \text{for all } s, t \subseteq S \text{ such that } s \cap t = \emptyset. \tag{10.5}$$

This rule is motivated by the idea that if this constraint were violated, the auctioneer could choose to ignore the bid $b_{i,s\cup t}$, and instead accept the bids $b_{i,s}$ and $b_{i,t}$. Thus, bidders can express preferences for complements, but their bids cannot be less for a combination than for the component parts. Cantillon and Pesendorfer (2003) extend the analysis to incorporate these constraints, showing that even in their presence, it is possible to place an upper bound on the extent of the synergies that exist between items.

11. Concluding remarks

The prominent role of auctions in allocating a wide range of public and private resources provides one strong motivation for empirical work on auctions. Recent methodological advances have made it possible to address old market design questions (e.g., how

to auction Treasury bills), while new policy questions (e.g., how to auction multiple complementary goods) have motivated development of new methodological tools. In addition, auctions hold the promise of shedding light on fundamental questions about the nature of information, preferences, and behavior that are of importance to a much broader scope of economic environments. Like earlier descriptive empirical work on auctions that provided influential evidence on the importance of asymmetric information and strategic behavior, recent empirical work using structural econometric models has also begun to deliver on this promise, addressing such questions as the empirical importance of reputations, entry costs, or adverse selection. Because of the close match between the theory and actual institutions, auctions have the potential to provide insights into fundamental questions that are difficult or impossible to address without the benefit of structure from economic theory. We expect much of the most interesting future work in the empirical auction literature to push farther in this direction.

It is worth noting that the analysis of identification in auction models is useful outside of the realm of econometrics. For example, in some models of learning in games, a central component of the analysis concerns whether it is possible to infer primitives of the game from the distribution of equilibrium outcomes that can be observed by players. The equilibrium concept of self-confirming equilibrium [Fudenberg and Levine (1993), Dekel, Fudenberg and Levine (2003)], motivated by learning models, hinges on just this issue.[117] Recently, Esponda (2004) analyzed self-confirming equilibria in auction games, focusing on the extent to which information revealed by an auctioneer allows bidders to infer the distribution over opponent types. This problem is closely related to the identification problem.[118]

Auctions have long been recognized as providing ideal market institutions for exploring the relationships between economic theory and the actual behavior of economic agents. Since the seminal work of Vickrey (1961) and Wilson (1967), rich theoretical and empirical literatures on auctions have developed. In our view, one of the most exciting advances in this literature is the development of methods for combining theoretical and statistical analysis in order to learn about the primitive features of an auction environment from observed bidding behavior. We have focused our discussion on nonparametric identification, in part because this makes transparent how the relationships derived from theory can be used to make valid inferences from data. We hope that this chapter will be a valuable reference and starting point for researchers who will apply and expand upon these methods to explore the wide range of open questions in the future.

[117] This concept relaxes the common knowledge assumption of Nash equilibrium, but requires that bidders best-respond to beliefs that are consistent with the equilibrium distribution of outcomes that is observable to the bidders. For example, the bidders might observe the distribution of transactions prices, or the distribution of all bids.

[118] Furthermore, this alternative to the standard common knowledge assumption may be an interesting possibility to explore in an empirical model.

References

Ackerberg, D., Benkard, C.L., Berry, S., Pakes, A. (2007). "Econometric tools for analyzing market outcomes". In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6A. Elsevier (Chapter 63).

Andersen, P.K., Borgan, Ø., Gill, R., Keiding, N. (1991). Statistical Models Based on Counting Processes. Springer, New York.

Aoyagi, M. (2003). "Bid rotation and collusion in repeated auctions". Journal of Economic Theory 112, 79–105.

Armantier, O., Sbaï, E. (2003). "Estimation and comparison of treasury auction formats when bidders are asymmetric". Working Paper. SUNY–Stony Brook.

Armstrong, M., Rochet, J.C. (1999). "Multidimensional screening: A user's guide". European Economic Review 43, 959–979.

Arnold, B., Balakrishnan, N., Nagaraja, H. (1992). A First Course in Order Statistics. Wiley & Sons, New York.

Athey, S. (2001). "Single crossing properties and the existence of pure strategy equilibrium in games of incomplete information". Econometrica 69, 861–890.

Athey, S., Bagwell, K. (2001). "Optimal collusion with private information". RAND Journal of Economics 32, 428–465.

Athey, S., Bagwell, K. (2004a). "Collusion with persistent cost shocks". Working Paper. Stanford University.

Athey, S., Bagwell, K. (2004b). "Dynamic auctions with persistent cost shocks". Working Paper. Stanford University.

Athey, S., Haile, P. (2000). "Identification of standard auction models". MIT Working Paper 00-18.

Athey, S., Haile, P. (2002). "Identification of standard auction models". Econometrica 70, 2107–2140.

Athey, S., Levin, J. (2001). "Information and competition in US Forest Service timber auctions". Journal of Political Economy 109, 375–417.

Athey, S., Bagwell, K., Sanchirico, C. (2004). "Collusion and price rigidity". The Review of Economic Studies 71, 317–349.

Athey, S., Levin, J., Seira, E. (2004). "Comparing open and sealed bid auctions: Theory and evidence from timber auctions". Working Paper. Stanford University.

Ausubel, L., Cramton, P. (2002). "Demand reduction and inefficiency in multi-unit auctions". Working Paper 96–07. University of Maryland.

Avery, C. (1998). "Strategic jump bidding in English auctions". Review of Economic Studies 65, 185–210.

Back, K., Zender, J. (1993). "Auctions of divisible goods: On the rationale for the US treasury experiment". Review of Financial Studies 6, 733–764.

Bajari, P. (1997). "The first-price auction with asymmetric bidders: Theory and applications". PhD Dissertation. University of Minnesota.

Bajari, P. (2001). "Comparing competition and collusion: A numerical approach". Economic Theory 18, 187–205.

Bajari, P., Hortaçsu, A. (2003a). "The winner's curse, reserve prices, and endogenous entry: Empirical insights from eBay auctions". RAND Journal of Economics 34, 329–355.

Bajari, P., Hortaçsu, A. (2003b). "Cyberspace auctions and pricing issues: A review of empirical findings". In: Jones, D. (Ed.), New Economy Handbook. Elsevier.

Bajari, P., Hortaçsu, A. (2004). "Economic insights from Internet auctions". Journal of Economic Literature 17, 457–486.

Bajari, P., Hortaçsu, A. (2005). "Are structural estimates of auction models reasonable? Evidence from experimental data". Journal of Political Economy 113, 703–741.

Bajari, P., Summers, G. (2002). "Detecting collusion in procurement auctions". Antitrust Law Journal 70, 143–170.

Bajari, P., Ye, L. (2003). "Deciding between competition and collusion". Review of Economics and Statistics 85, 971–989.

Bajari, P., Houghton, S., Tadelis, S. (2004). "Bidding for incomplete contracts". Working Paper. Duke University.

Balasubramanian, K., Balakrishnan, N. (1994). "Equivalence of relations for order statistics for exchangeable and arbitrary cases". Statistics and Probability Letters 21, 405–407.
Baldwin, L., Marshall, R., Richard, J. (1997). "Bidder collusion in US Forest Service timber sales". Journal of Political Economy 105, 657–699.
Bartholomew, D. (1959). "A test of homogeneity for ordered alternatives". Biometrika 46, 36–48.
Berman, S. (1963). "Note on extreme values, competing risks, and semi-Markov processes". Annals of Mathematical Statistics 34, 1104–1106.
Berry, S., Reiss, P. (in press). "Empirical models of entry and market structure". In: Armstrong, M., Porter, R. (Eds.). Handbook of Industrial Organization, vol. III. Elsevier.
Berry, S., Tamer, E. (2005). "Identification in models of oligopoly entry". Working Paper. Yale University.
Berry, S., Levinsohn, J., Pakes, A. (1995). "Automobile prices in market equilibrium". Econometrica 63, 841–890.
Bikhchandani, S. (1988). "Reputation in repeated second-price auctions". Journal of Economic Theory 46, 97–119.
Bikhchandani, S., Huang, C. (1989). "Auctions with resale markets: An exploratory model of treasury bill markets". Review of Financial Studies 2, 311–339.
Bikhchandani, S., Haile, P., Riley, J. (2002). "Symmetric separating equilibria in English auctions". Games and Economic Behavior 38, 19–27.
Borenstein, S., Bushnell, J., Wolak, F. (2002). "Measuring market inefficiencies in California's restructured wholesale electricity market". American Economic Review 92, 1376–1405.
Bowman, A., Jones, M., Gijbels, I. (1998). "Testing monotonicity of a regression". Journal of Computational and Graphical Statistics 7, 489–500.
Brannman, L., Klein, D., Weiss, L. (1987). "The price effects of increased competition in auction markets". Review of Economics and Statistics 69, 24–32.
Brendstrup, B., Paarsch, H. (2003). "Nonparametric estimation of Dutch and first-price, sealed-bid auction models with asymmetric bidders". Working Paper. University of Iowa.
Brendstrup, B., Paarsch, H. (2004). "Nonparametric identification and estimation of multi-unit, sequential, oral ascending-price auctions with asymmetric bidders". Working Paper. University of Iowa.
Campo, S. (2002). "Asymmetry and risk aversion within the independent private values paradigm: The case of construction procurement contracts". Working Paper. University of Southern California.
Campo, S., Guerre, E., Perrigne, I., Vuong, Q. (2002). "Semiparametric estimation of first-price auctions with risk averse bidders". Working Paper. Pennsylvania State University.
Campo, S., Perrigne, I., Vuong, Q. (2003). "Asymmetry in first-price auctions with affiliated private values". Journal of Applied Econometrics 18, 197–207.
Cantillon, E., Pesendorfer, M. (2003). "Combination bidding in multi-unit auctions". Working Paper. London School of Economics and Political Science.
Chen, X. (2007). "Large sample sieve estimation of semi-nonparametric models". In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier (Chapter 76).
Chernozhukov, V., Hong, H. (2003). "Likelihood inference in a class of nonregular econometric models". Econometrica 72, 1445–1480.
Chow, Y., Teicher, H. (1997). Probability Theory: Independence, Interchangeability and Martingales. Springer, New York.
Crémer, J., McLean, R. (1988). "Full extraction of the surplus in Bayesian and dominant strategy auctions". Econometrica 56, 1247–1257.
David, H. (1981). Order Statistics. Wiley, New York.
David, H., Moeschberger, M. (1978). The Theory of Competing Risks. Macmillan, New York.
Das Varma, G. (2003). "Bidding for a process innovation under alternative modes of competition". International Journal of Industrial Organization.
Dekel, E., Fudenberg, D., Levine, D. (2003). "Learning to play Bayesian games". Games and Economic Behavior 46, 282–303.
Donald, S., Paarsch, H. (1993). "Piecewise pseudo-maximum likelihood estimation in empirical models of auctions". International Economic Review 34, 121–148.

Donald, S., Paarsch, H. (1996). "Identification, estimation, and testing in parametric empirical models of auctions within the independent private values paradigm". Econometric Theory 12, 517–567.

Donald, S., Paarsch, H., Robert, J. (2006). "An empirical model of the multi-unit, sequential clock auction". Journal of Applied Econometrics 21, 1221–1247.

Einav, L. (2004). "A note on the analogies between empirical models of auctions and of differentiated product markets". Working Paper. Stanford University.

Esponda, I. (2004). "Information feedback and self-confirming equilibrium in first price auctions". Working Paper. Stanford University.

Fermanian, J. (2003). "Nonparametric estimation of competing risks models with covariates". Journal of Multivariate Analysis 85, 156–191.

Février, P. (2004). "Semiparametric identification and estimation of common value auctions". Working Paper. CREST.

Février, P., Préget, R., Visser, M. (2002). "Econometrics of share auctions". Working Paper. CREST.

Flambard, V., Perrigne, I. (2006). "Asymmetry in procurement auctions: Some evidence from snow removal contracts". Economic Journal, Royal Economic Society 116 (514), 1014–1036.

Friedman, M. (1960). A Program for Monetary Stability. Fordham University Press, New York.

Fudenberg, D., Levine, D. (1993). "Self-confirming equilibrium". Econometrica 61, 523–546.

Fudenberg, D., Tirole, J. (1991). Game Theory. MIT Press, Cambridge.

Gallant, R., Nychka, D. (1987). "Semi-nonparametric maximum likelihood estimation". Econometrica 55, 363–390.

Gijbels, I., Hall, P., Jones, M.C., Koch, I. (2000). "Tests for monotonicity of a regression mean with guaranteed level". Biometrika 87, 663–673.

Gilley, O., Karels, G. (1981). "The competitive effect in bonus bidding: New evidence". Bell Journal of Economics 12, 637–648.

Goeree, J. (2003). "Bidding for the future: Signaling in auctions with an aftermarket". Journal of Economic Theory 108, 345–364.

Guerre, E., Perrigne, I., Vuong, Q. (1995). "Nonparametric estimation of first-price auctions". Working Paper #9504. University of Southern California.

Guerre, E., Perrigne, I., Vuong, Q. (2000). "Optimal nonparametric estimation of first-price auctions". Econometrica 68, 525–574.

Haile, P. (1999). "Auctions with resale". Working Paper. University of Wisconsin.

Haile, P. (2001). "Auctions with resale markets: An application to US Forest Service timber sales". American Economic Review 91, 399–427.

Haile, P. (2003). "Auctions with private uncertainty and resale opportunities". Journal of Economic Theory 108, 72–110.

Haile, P., Tamer, E. (2001). "Inference from English auctions with asymmetric affiliated private values". Working Paper. University of Wisconsin-Madison.

Haile, P., Tamer, E. (2002). "Inference with an incomplete model of English auctions". Working Paper. Princeton University.

Haile, P., Tamer, E. (2003). "Inference with an incomplete model of English auctions". Journal of Political Economy 111, 1–52.

Haile, P., Hong, H., Shum, M. (2003). "Nonparametric tests for common values in first-price sealed-bid auctions". NBER Working Paper 10105.

Hall, P., Heckman, N. (2000). "Testing for monotonicity of a regression mean by calibrating for linear functions". Annals of Statistics 28, 20–39.

Harstad, R., Rothkopf, M. (2000). "An 'Alternating Recognition' model of English auctions". Management Science 46, 1–12.

Heckman, J., Honoré, B. (1989). "The identifiability of the competing risks model". Biometrika 76, 325–330.

Heckman, J., Honoré, B. (1990). "The empirical content of the Roy model". Econometrica 58, 1121–1149.

Hendricks, K., Paarsch, H. (1995). "A survey of recent empirical work concerning auctions". Canadian Journal of Economics 28, 403–426.

Hendricks, K., Porter, R. (1988). "An empirical study of an auction with asymmetric information". American Economic Review 78, 865–883.
Hendricks, K., Porter, R. (in press). "Lectures on auctions: An empirical perspective". In: Armstrong, M., Porter, R. (Eds.). Handbook of Industrial Organization, vol. III. Elsevier.
Hendricks, K., Porter, R., Boudreau, B. (1987). "Information, returns, and bidding behavior in OCS auctions: 1954–1969". Journal of Industrial Economics 35, 517–542.
Hendricks, K., Porter, R., Wilson, C. (1994). "Auctions for oil and gas leases with an informed bidder and a random reservation price". Econometrica 62, 1415–1444.
Hendricks, K., Pinkse, J., Porter, R. (2003). "Empirical implications of equilibrium bidding in first-price, symmetric, common value auctions". Review of Economic Studies 70, 115–145.
Hirano, K., Porter, J. (2003). "Asymptotic efficiency in parametric structural models with parameter-dependent support". Econometrica 71, 1307–1338.
Hollander, M., Wolfe, D. (1999). Nonparametric Statistical Methods. John Wiley and Sons, New York.
Hong, H., Paarsch, H.J. (2006). An Introduction to the Econometrics of Auction Data. MIT Press, Cambridge.
Hong, H., Shum, M. (2000). "Structural estimation of auction models". In: Patrone, F., Garcia-Jurado, I., Tijs, S. (Eds.), Game Practice: Contributions from Applied Game Theory. Kluwer, Boston.
Hong, H., Shum, M. (2002). "Increasing competition and the winner's curse: Evidence from procurement". Review of Economic Studies 69, 871–898.
Hong, H., Shum, M. (2003). "Econometric models of ascending auctions". Journal of Econometrics 112, 327–358.
Hortaçsu, A. (2002). "Mechanism choice and strategic bidding in divisible good auctions: An empirical analysis of the Turkish treasury auction market". Working Paper. University of Chicago.
Hotz, J., Miller, R. (1993). "Conditional choice probabilities and the estimation of dynamic models". Review of Economic Studies 60, 497–529.
Izmalkov, S. (2003). "English auctions with reentry". Working Paper. MIT.
Jackson, M., Swinkels, J. (2005). "Existence of equilibria in single and double private value auctions". Econometrica 73 (1), 93–140.
Jackson, M., Simon, L., Swinkels, J., Zame, W. (2002). "Equilibrium, communication, and endogenous sharing rules in discontinuous games of incomplete information". Econometrica 70, 1711–1740.
Jofre-Bonet, M., Pesendorfer, M. (2000). "Bidding behavior in repeated procurement auctions". European Economic Review 44, 1006–1020.
Jofre-Bonet, M., Pesendorfer, M. (2003). "Estimation of a dynamic auction game". Econometrica 71, 1443–1489.
Judd, K. (1998). Numerical Methods in Economics. MIT Press, Cambridge.
Kagel, J. (1995). "Auctions: A survey of experimental research". In: Kagel, J., Roth, A. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, pp. 501–585.
Kastl, J. (2005). "Discrete bids and empirical inference in divisible good auctions. Working Paper. Northwestern University.
Katzman, B., Rhodes-Kropf, M. (2002). "The consequences of information revealed in auctions". Working Paper. Columbia University.
Klemperer, P. (1999). "Auction theory: A guide to the literature". Journal of Economic Surveys 13, 227–286.
Klemperer, P. (2002). "What really matters in auction design". Journal of Economic Perspectives 16, 169–189.
Koopmans, T. (1945). "Statistical estimation of simultaneous economic relations". Journal of the American Statistical Association 40, 448–466.
Kotlarski, I. (1966). "On some characterization of probability distributions in Hilbert spaces". Annali di Matematica Pura ed Applicata 74, 129–134.
Krasnokutskaya, E. (2004). "Auction models with unobserved heterogeneity: Application to the Michigan highway procurement auctions". Working Paper. University of Pennsylvania.
Krishna, V. (2002). Auction Theory. Academic Press, San Diego.
Laffont, J. (1997). "Game theory and empirical economics: The case of auction data". European Economic Review 41, 1–35.

Laffont, J., Vuong, Q. (1993). "Structural econometric analysis of descending auctions". European Economic Review 37, 329–341.
Laffont, J., Vuong, Q. (1996). "Structural analysis of auction data". American Economic Review, Papers and Proceedings 86, 414–420.
Laffont, J., Ossard, H., Vuong, Q. (1995). "Econometrics of first-price auctions". Econometrica 63, 953–980.
Lancaster, T. (1990). The Econometric Analysis of Transition Data: An Econometric Society Monograph. Cambridge University Press.
Lebrun, B. (1999). "First price auctions in the asymmetric N bidder case". International Economic Review 40, 125–142.
Levin, D., Smith, J. (1994). "Equilibrium in auctions with entry". American Economic Review 84, 585–599.
Li, T. (2002). "Robust and consistent estimation of nonlinear errors-in-variables models". Journal of Econometrics 110, 1–26.
Li, T. (2003). "Econometrics of first-price auctions with entry and binding reservation prices". Working Paper. Indiana University.
Li, T., Vuong, Q. (1998). "Nonparametric estimation of the measurement error model using multiple indicators". Journal of Multivariate Analysis 65, 139–165.
Li, T., Zheng, X. (2005). "Procurement auctions with entry and an uncertain number of actual bidders: Theory, structural inference, and an application". Working Paper. Indiana University.
Li, T., Perrigne, I., Vuong, Q. (2000). "Conditionally independent private information in OCS wildcat auctions". Journal of Econometrics 98, 129–161.
Li, T., Perrigne, I., Vuong, Q. (2002). "Structural estimation of the affiliated private value auction model". RAND Journal of Economics 33, 171–193.
Lizzeri, A., Persico, N. (2000). "Uniqueness and existence of equilibrium in auctions with a reserve price". Games and Economic Behavior 30, 83–114.
Loader, C. (1999). Local Regression and Likelihood. Springer, New York.
Lucking-Reiley, D. (2000). "Auctions on the Internet: What's being auctioned, and how?". Journal of Industrial Economics 48, 227–252.
Manski, C. (1995). Identification Problems in the Social Sciences. Harvard University Press.
Manski, C., Tamer, E. (2002). "Inference on regressions with interval data on a regressor or outcome". Econometrica 70, 519–546.
Maskin, E., Riley, J. (1984). "Optimal auctions with risk averse buyers". Econometrica 52, 1473–1518.
Maskin, E., Riley, J. (2000a). "Asymmetric auctions". Review of Economic Studies 67, 413–438.
Maskin, E., Riley, J. (2000b). "Equilibrium in sealed high bid auctions". Review of Economic Studies 67, 439–454.
Maskin, E., Riley, J. (2003). "Uniqueness of equilibrium in sealed high-bid auctions". Games and Economic Behavior 45, 395–409.
Matthews, S. (1984). "Information acquisition in discriminatory auctions". In: Boyer, M., Kihlstrom, R.E. (Eds.), Bayesian Models in Economic Theory. North-Holland, New York.
McAdams, D. (2004a). "Monotone equilibrium in multi-unit auctions". Working Paper. MIT.
McAdams, D. (2004b). "Uniqueness in first-price auctions with affiliation". Working Paper. MIT.
McAdams, D. (2005). "Identification and testable restrictions in private value multi-unit auctions". Working Paper. MIT.
McAdams, D. (2007). "Monotonicity in asymmetric first-price auctions with affiliation". International Journal of Game Theory 35, 427–453.
McAfee, R.P., McMillan, J. (1987). "Auctions and bidding". Journal of Economic Literature 25, 669–738.
McAfee, R.P., McMillan, J. (1992). "Bidding rings". American Economic Review 82, 579–599.
McAfee, R.P., McMillan, J. (1996). "Analyzing the airwaves auction". Journal of Economic Perspectives 10, 159–175.
McAfee, R.P., Reny, P. (1992). "Correlated information and mechanism design". Econometrica 60, 395–421.
McAfee, R.P., McMillan, J., Whinston, M. (1989). "Multiproduct monopoly, commodity bundling and correlation of values". Quarterly Journal of Economics 102, 371–383.

McAfee, R.P., Quan, D., Vincent, D. (2002). "Minimum acceptable bids, with application to real estate auctions". Journal of Industrial Economics 50, 391–416.
McAfee, R.P., Takacs, W., Vincent, D.R. (1999). "Tariffying auctions". RAND Journal of Economics 30, 158–179.
McFadden, D. (1989). "Testing for stochastic dominance". In: Studies in the Economics of Uncertainty: In Honor of Josef Hadar. Springer, New York.
Meilijson, I. (1981). "Estimation of the lifetime distribution of the parts from the autopsy statistics of the machine". Journal of Applied Probability 18, 829–838.
Milgrom, P. (1981). "Good news and bad news: Representation theorems and applications". Bell Journal of Economics 12, 380–391.
Milgrom, P. (2004). Putting Auction Theory to Work. Cambridge University Press, Cambridge.
Milgrom, P.R., Weber, R.J. (1982). "A theory of auctions and competitive bidding". Econometrica 50, 1089–1122.
Myerson, R. (1981). "Optimal auction design". Mathematics of Operations Research 6, 58–73.
Ockenfels, A., Roth, A.E. (2006). "Late and multiple bidding in second-price Internet auctions: Theory and evidence concerning different rules for ending an auction". Games and Economic Behavior 55 (2), 297–320.
Olley, G.S., Pakes, A. (1996). "The dynamics of productivity in the telecommunications equipment industry". Econometrica 64, 1263–1297.
Paarsch, H. (1992a). "Deciding between common and private values paradigms in empirical models of auctions". Journal of Econometrics 51, 191–215.
Paarsch, H. (1992b). "Empirical models of auctions and an application to British Columbian timber sales". Research Report 9212. University of Western Ontario.
Paarsch, H. (1994). "A comparison of estimators for empirical models of auctions". Annales d'Economie et de Statistique 34, 143–157.
Paarsch, H. (1997). "Deriving an estimate of the optimal reserve price: An application to British Columbian timber sales". Journal of Econometrics 78, 333–357.
Perrigne, I. (2003). "Random reserve prices and risk aversion in timber sale auctions". Working Paper. Pennsylvania State University.
Perrigne, I., Vuong, Q. (1999). "Structural econometrics of first-price auctions: A survey of methods". Canadian Journal of Agricultural Economics 47, 203–223.
Pesendorfer, M. (2000). "A study of collusion in first-price auctions". Review of Economic Studies 67, 381–411.
Peters, M., Severinov, S. (2006). "Internet auctions with many traders". Journal of Economic Theory 130, 220–245.
Pinkse, J., Tan, G. (2005). "The affiliation effect in first-price auctions". Econometrica 73, 263–277.
Politis, D., Romano, J., Wolf, M. (1999). Subsampling. Springer-Verlag, New York.
Porter, R., Zona, J.D. (1993). "Detection of bid rigging in procurement auctions". Journal of Political Economy 101, 518–538.
Porter, R., Zona, J.D. (1999). "Ohio school milk markets: An analysis of bidding". RAND Journal of Economics 30, 263–288.
Prakasa-Rao, B.L.S. (1992). Identifiability in Stochastic Models: Characterization of Probability Distributions. Academic Press, San Diego.
Quint, D. (2004). "Optimal second price auctions with positively correlated private values and limited information". SIEPR Discussion Paper 03-14. Stanford University.
Reiss, P., Wolak, F. (2007). "Structural econometric modeling: Rationales and examples from industrial organization". In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6A. Elsevier (Chapter 64).
Reny, P., Zamir, S. (2004). "On the existence of pure strategy monotone equilibria in asymmetric first-price auctions". Econometrica 72, 1105–1125.
Romano, J. (1988). "A bootstrap revival of some nonparametric distance tests". Journal of the American Statistical Association 83, 698–708.

Romano, J. (1989). "Bootstrap and randomization tests of some nonparametric hypotheses". Annals of Statistics 17, 141–159.
Samuelson, W. (1985). "Competitive bidding with entry costs". Economics Letters 17, 53–57.
Schennach, S. (2004). "Estimation of nonlinear models with measurement error". Econometrica 72, 33–75.
Shneyerov, A. (2005). "An empirical study of auction revenue rankings: The case of municipal bonds". Working Paper. University of British Columbia.
Silverman, B. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
Skryzpacz, A., Hopenhayn, H. (2004). "Tacit collusion in repeated auctions". Journal of Economic Theory 114, 153–169.
Smiley, A. (1979). Competitive Bidding Under Uncertainty: The Case of Offshore Oil. Ballinger, Cambridge.
Song, U. (2003). "Nonparametric estimation of an eBay auction model with an unknown number of bidders". Working Paper. University of Wisconsin.
Song, U. (2004). "Structural analysis of auction data with an unknown number of bidders". PhD Dissertation. University of Wisconsin.
Su, L., White, H. (2003). "Testing conditional independence via empirical likelihood". Working Paper. University of California San Diego.
Thiel, S. (1988). "Some evidence on the winner's curse". American Economic Review 78, 884–895.
Wang, M., Jewell, N.P., Tsai, W. (1986). "Asymptotic properties of the product limit estimate under random truncation". Annals of Statistics 14, 1597–1605.
Whinston, M. (1989). "Tying, foreclosure, and exclusion". American Economic Review 90, 837–859.
Wilson, R. (1967). "Competitive bidding with asymmetric information". Management Science 11, 816–820.
Wilson, R. (1979). "Auctions of shares". Quarterly Journal of Economics 93, 675–689.
Wolak, F. (2003). "Identification and estimation of cost functions using observed bid data: An application to electricity markets". In: Dewatripont, M., Hansen, L., Turnovsky, P. (Eds.), Advances in Economics and Econometrics – Theory and Applications. Eighth World Congress. In: Econometric Society Monographs, vol. 2. Cambridge University Press, Cambridge, pp. 115–149.
Wolak, F. (2004). "Quantifying the supply-side benefits from forward contracting in wholesale electricity markets. Working Paper. Stanford University.
Wolfram, C. (1998). "Strategic bidding in a multi-unit auction, an empirical analysis of bids to supply electricity in England and Wales". RAND Journal of Economics 29, 703–725.
Woodroofe, M. (1985). "Estimating a distribution function with truncated data". Annals of Statistics 13, 163–177.
Vickrey, W. (1961). "Counterspeculation, auctions, and competitive sealed tenders". Journal of Finance 16, 8–37.
Yin, P. (2004). "eBay auctions as markets". Working Paper. Harvard Business School.
Zulehner, C. (2003). "Bidding behavior and bidders' valuations in Austrian cattle auctions". Working Paper. University of Vienna.

Chapter 61

INTERTEMPORAL SUBSTITUTION AND RISK AVERSION*

LARS PETER HANSEN, JOHN HEATON, JUNGHOON LEE and NIKOLAI ROUSSANOV

University of Chicago, USA
e-mails: l-hansen@uchicago.edu; john.heaton@gsb.uchicago.edu; nroussan@gsb.uchicago.edu; junghoon@uchicago.edu

Contents

* Nan Li provided expert research assistance. We benefitted from comments by Nan Li, Sydney Ludvigson and Jesus Santos. Hansen acknowledges support from the National Science Foundation under Award Number SES0519372. Heaton acknowledges support from the Center for Research in Security Prices.

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06061-8

Abstract

We study structural models of stochastic discount factors and explore alternative methods of estimating such models using data on macroeconomic risk and asset returns.

Particular attention is devoted to recursive utility models in which risk aversion can be modified without altering intertemporal substitution. We characterize the impact of changing the intertemporal substitution and risk aversion parameters on equilibrium short-run and long-run risk prices and on equilibrium wealth.

Keywords

intertemporal substitution, risk aversion, recursive utility, GMM, asset pricing

JEL classification: C1, C32, E21, G12

1. Introduction

Households save and invest both for intertemporal reasons and to control exposure to risk. The resulting patterns of consumption, savings and investment, at both the household and the aggregate level, reveal information about the parameters of preferences that govern intertemporal substitution and risk aversion. Prices that clear financial markets must also reflect the demands of investors and hence are affected by their preferences. In this way security market data convey information from asset prices that complements that from microeconomic data sets, from experimental evidence, or from survey evidence. An important aim of this chapter is to understand better how changes in investor preferences alter asset prices. This guides our understanding of the consequences of inputs from external data sources and the value of asset market data for revealing investor preferences.

Risk premia in security returns provide compensation for risk averse investors. These risk premia often have simple characterizations. For instance, in the capital asset pricing model (CAPM), risk premia are proportional to the covariances between the return to the aggregate wealth portfolio and asset returns. More generally, in the consumption-based capital asset pricing model (CCAPM) the covariance between consumption and asset returns determines the riskiness of returns. Since the dynamics of consumption are linked to the dynamics of wealth, this model implies that understanding the riskiness of the wealth of investors is ultimately important in understanding security returns. This linkage is heavily influenced by the assumed form of investor preferences.

While asset market data offer fertile proving ground for theories of investor behavior and market structure, this data source also poses special challenges or puzzles. In the case of the CAPM, differences across securities in the measured covariance of returns with aggregate stock market indices have been shown to have little relationship with average returns [see for example Fama and French (1992)]. Similarly there appears to be very little covariance between measures of the aggregate consumption of investors, and asset returns. The empirical results in Grossman and Shiller (1981), Hansen and Singleton (1983), Mehra and Prescott (1985), Shiller (1982) and Hansen and Jagannathan (1991) give alternative characterizations of *puzzles* that emerge in the study of security market returns and aggregate consumption. Thus, when we look to security market data for information about preference parameters, we are exposed to the empirical challenges posed by this source of data.

Our chapter features alternative and complementary methods of analysis for the study of the macroeconomic underpinning of asset valuation. We describe some alternative ways to characterize model implications, and we show how statistical methods can be put to good use. While we apply some of these methods to illustrate substantive results, our chapter is not intended as comprehensive of empirical evidence. Excellent surveys with more extensive empirical discussions are given in Campbell (2003) and Lettau and Ludvigson (2003).

Alternative specifications of investor preferences and their links to prices are given in Sections 2 and 3. Specifically in Section 3 we show how to construct stochastic discount

factors used in representing prices for the alternative models of investor preferences described in Section 2. While we describe the investor preferences for an array of models, we focus our equilibrium price calculations and quantification on a particular subset of these preferences, the CES Kreps and Porteus (1978) model. This model is rich enough to draw an interesting distinction between risk aversion and intertemporal substitution and to pose important measurement and econometric challenges. Some basic statistical methods for characterizing present-value implications as they relate to asset pricing are developed in Section 4. Section 5 develops some analytical results and local approximations designed to reveal how intertemporal substitution and risk aversion alter equilibrium prices. Section 6 uses vector-autoregressive (VAR) statistical models to measure risk aversion from a heterogenous set of asset returns and quantifies the resulting statistical uncertainty. Section 7 develops generalized method of moments (GMM) and related estimation methods and illustrates their use in extracting measures of intertemporal substitution and risk aversion. These latter sections add some important qualifications to the existing empirical literature.

2. Investor preferences

In this section we survey a variety of models of investor preferences that are used in the literature. These specifications of investor preferences imply, through their intertemporal marginal rates of substitution, stochastic discount factors that represent risk prices and interest rates. This discussion is complementary to the Backus, Routledge and Zin (2004) survey of *exotic preferences* pertinent to macroeconomics. As in what follows, they provide alternative specifications of intertemporal and risk preferences.[1]

Recursive utility gives a useful framework for parameterizing risk aversion and intertemporal substitution. As advocated by Epstein and Zin (1989a) and Constantinides (1990), it gives a convenient way to introduce parameters that serve distinct roles in decision making. Let $\{\mathcal{F}_t\colon\ t \geqslant 0\}$ denote the sequence of conditioning information sets (sigma algebras) available to an investor at dates $t = 0, 1, \ldots$. Adapted to this sequence are consumption processes $\{C_t\colon\ t \geqslant 0\}$ and a corresponding sequence of continuation values $\{V_t\colon\ t \geqslant 0\}$ associated with this consumption process. The date t components C_t and V_t are restricted to be in the date t conditioning information set.[2] The continuation values are determined recursively and used to rank alternative consumption processes.

Consider three approaches. The first approach takes a risk adjustment of the continuation value; the second approach introduces intertemporal complementarities; and the third approach social externalities.

[1] While Backus, Routledge and Zin (2004) do an admirable job of describing a broad class of preference specifications and their use in macroeconomics, the empirical challenge is how to distinguish among these alternatives. As Hansen (2004) emphasizes, some specifications are inherently very difficult to distinguish from one another.

[2] More formally, C_t and V_t are restricted to be $\mathcal{F}_t$ measurable.

2.1. Risk adjustment

Consider investor preferences that can be represented recursively as

$$V_t = \psi(C_t, V_{t+1}|\mathcal{F}_t)$$

where C_t is current consumption. Given a consumption process, this recursion takes future values and maps them into current values. It requires a terminal condition for the continuation value to initiate a backward induction. A familiar example is:

$$V_t = (1-\beta)U(C_t) + \beta E(V_{t+1}|\mathcal{F}_t)$$

where U is a concave utility function. This recursion is additive in *expected utility*. More general depictions of recursive utility provide a way to allow for alternative adjustments to risk and uncertainty.

2.1.1. A smooth adjustment

Following Kreps and Porteus (1978) and Epstein and Zin (1989a), introduce a strictly increasing, smooth concave function h. In applications this function is typically

$$h(V) = \begin{cases} \frac{V^{1-\gamma}-1}{1-\gamma}, & \gamma > 0, \gamma \neq 1, \\ \log V, & \gamma = 1. \end{cases}$$

Then a risk adjusted version of the continuation value is

$$\mathsf{R}(V|\mathcal{F}) \doteq h^{-1}\big(E\big[h(V)|\mathcal{F}\big]\big).$$

The presumption is that V_t depends on the continuation value through the risk adjustment $\mathsf{R}(V_{t+1}|\mathcal{F}_t)$, which is a restriction on function ψ:

$$V_t = \psi(C_t, V_{t+1}|\mathcal{F}_t) = \psi^*\big[C_t, \mathsf{R}(V_{t+1}|\mathcal{F}_t)\big].$$

The function h is strictly increasing and adjusts for the riskiness of the continuation value for the consumption profile $\{C_{t+\tau}\colon\ \tau = 1, 2, \ldots\}$. It imposes a nontrivial preference over lotteries indexed by calendar time. The parametric form of h gives a convenient way to parameterize risk preferences.

Consider the special case in which the continuation value is perfectly predictable, implying that $E(V_{t+1}|\mathcal{F}_t) = V_{t+1}$. Then $\mathsf{R}(V_{t+1}|\mathcal{F}_t) = V_{t+1}$ so that the function h has no bearing on the specification of preferences over perfectly forecastable consumption plans. The incremental risk adjustment does alter the implications for intertemporal substitution for predictable consumption plans.

Examples of ψ^* function are as follows:

EXAMPLE 2.1.

$$\psi^*(C, \mathsf{R}) = (1-\beta)U(C) + \beta\mathsf{R}$$

for some increasing concave function U.

The concavity of U already induces some degree of risk aversion, but it also has an impact on intertemporal substitution.

It is often convenient to work with an aggregator that is homogeneous of degree one. Curvature in U can be offset by transforming the continuation value. In the case of a constant elasticity of substitution (CES) specification this gives:

EXAMPLE 2.2.

$$\psi^*(C,\mathsf{R}) = \left[(1-\beta)(C)^{1-\rho} + \beta(\mathsf{R})^{1-\rho}\right]^{\frac{1}{1-\rho}}$$

for $\rho > 0$. The case in which $\rho = 1$ requires taking limits, and results in the Cobb–Douglas specification:

$$\psi^*(C,\mathsf{R}) = C^{1-\beta}\mathsf{R}^{\beta}.$$

The parameter ρ is the reciprocal of the elasticity of intertemporal substitution.

EXAMPLE 2.3. There is an extensive literature in control theory starting with the work of Jacobson (1973) and Whittle (1990) on introducing risk sensitivity into control problems. Hansen and Sargent (1995) suggest a recursive version of this specification in which

$$\psi^*(C,\mathsf{R}) = U(C) + \beta\mathsf{R}$$

as in Example 2.1 with the incremental risk adjustment given by

$$\mathsf{R}(V_{t+1}|\mathcal{F}_t) = -\frac{1}{\theta}\log E\left[\exp(-\theta V_{t+1})|\mathcal{F}_t\right].$$

The parameter θ is the risk sensitivity parameter. As emphasized by Tallarini (1998), this specification overlaps with the CES specification when $\rho = 1$, $U(C) = \log C$ and $\theta = \gamma - 1$. To verify this link, take logarithms of the continuation values in the CES recursions. The logarithmic function is increasing and hence ranks of hypothetical consumption processes are preserved.

Although it is convenient to make a risk adjustment of the continuation value, there is an alternative transformation of the continuation value that depicts preferences as a nonlinear version of expected utility. Let

$$\widetilde{V}_t = h(V_t).$$

Then

$$\widetilde{V}_t = h\left[\psi^*\left(C_t, h^{-1}\left[E(\widetilde{V}_{t+1}|\mathcal{F}_t)\right]\right)\right] = \tilde{\psi}\left[C_t, E(\widetilde{V}_{t+1}|\mathcal{F}_t)\right].$$

The introduction of h can induce nonlinearity in the aggregator $\tilde{\psi}$. Kreps and Porteus (1978) use such a nonlinear aggregator to express a preference for early and late resolution of uncertainty. When $\tilde{\psi}$ is convex in this argument there is a preference for early resolution of uncertainty and conversely when $\tilde{\psi}$ is concave. We will show that the intertemporal composition of risk also matters for asset pricing.

2.1.2. A version without smoothness

The Epstein and Zin (1989a) recursive formulation was designed to accommodate more fundamental departures from the standard expected utility model. This includes departures in which there are *kinks* in preferences inducing first-order risk aversion. First-order risk aversion is used in asset pricing as a device to enhance risk aversion.

Examples of applications in the asset pricing include Bekaert, Hodrick and Marshall (1997) and Epstein and Zin (1990), but we shall feature a more recent specification due to Routledge and Zin (2003). Routledge and Zin (2003) propose and motivate an extension of Gul (1991)'s preferences for disappointment aversion. These preferences are based on a different way to compute the risk adjustment to a continuation value and induce first-order risk aversion. Continuation values are risk adjusted in accordance to

$$h(\widetilde{V}) = E[h(V)|\mathcal{F}] + \alpha E\left(\mathbf{1}_{\{V-\delta\widetilde{V}\leqslant 0\}}\left[h(V) - h(\delta\widetilde{V})\right]|\mathcal{F}\right)$$

which is an implicit equation in $\widetilde{V}$. In this equation, **1** is used as the indicator function of the subscripted event. The random variable $h(\widetilde{V})$ is by construction less than or equal to the conditional expectation of $h(V)$ with an extra negative contribution coming because of the averaging over the bad events defined by the threshold $h(V) \leqslant h(\delta\widetilde{V})$. The risk adjusted value is defined to be

$$\mathsf{R}(V|\mathcal{F}) = \widetilde{V}.$$

The h function is used as a risk adjustment as in our previous construction, but the parameters $0 < \delta < 1$ and $\alpha > 0$ capture a notion of *disappointment aversion*. While the Gul (1991) specification assumes that $\delta = 1$, this limits the preference kink to be on the certainty line. By allowing δ to be less than one, the disappointment cutoff is allowed to be lower.

2.2. Robustness and uncertainty aversion

Investors may be unsure about the probability used to evaluate risks. Instead of using one model, they may choose a family of such models. In some circumstances this also leads to what looks like a risk adjustment in the continuation value to a consumption plan. We illustrate this using the well-known close relationship between *risk sensitivity* and *robustness* featured in the control theory literature starting with the work of Jacobson (1973). As in Hansen and Sargent (1995) we may formulate this recursively as

$$\begin{aligned} v_t = {} & (1-\beta)U(C_t) \\ & + \min_{q_{t+1}\geqslant 0,\, E(q_{t+1}|\mathcal{F}_t)=1} \left[\beta E(q_{t+1}v_{t+1}|\mathcal{F}_t) + \beta\theta E(q_{t+1}\log q_{t+1}|\mathcal{F}_t)\right] \end{aligned}$$

where θ is a penalization parameter and q_{t+1} is a random variable used to distort the conditional probability distribution. The minimization is an adjustment for uncertainty in the probability model, and $E[q_{t+1}(\log q_{t+1})|\mathcal{F}_t]$ is a discrepancy measure for the probability distortion called conditional relative entropy.

The solution to the minimization problem is to set

$$q_{t+1} \propto \exp\left(-\frac{v_{t+1}}{\theta}\right)$$

where the proportional constant is conditional on $\mathcal{F}_t$ and chosen so that $E(q_{t+1}|\mathcal{F}_t) = 1$. This solution gives an *exponential tilt* to the original conditional probability distribution based on the continuation value and penalty parameter θ. Substituting this minimized choice of q_{t+1} gives the recursion:

$$v_t = (1-\beta)U(C_t) + \beta h^{-1} E\big[h(v_{t+1})|\mathcal{F}_t\big] \tag{1}$$

where

$$h(v) = \exp(-v/\theta).$$

Hence this setting is equivalent to assuming an exponential risk adjustment in the continuation value function.

As emphasized by Tallarini (1998), when U is the logarithmic function, we may transform the continuation value of (1) to obtain the Cobb–Douglas recursion in Example 2.2 with $\theta = \frac{1}{\gamma - 1}$ and $V_t = \exp(v_t)$. Maenhout (2004) and Skiadas (2003) give a characterization of this link in more general circumstances that include the CES specification in a continuous time version of these preferences by making the penalization depend on the endogenous continuation value [see also Hansen (2004)].

Strictly speaking, to establish a formal link between inducing a concern about model misspecification and a concern about risk required a special set of assumptions. These results illustrate, however, that it may be difficult in practice to disentangle the two effects. What may appear to be risk aversion emanating from asset markets may instead be a concern that a conjectured or benchmark probability model is inaccurate. Risk aversion from asset market data may be different from risk aversion in an environment with well-defined probabilities.

There are other ways to model uncertainty aversion. Following Epstein and Schneider (2003) we may constrain the family of probabilities period by period instead penalizing deviations. If we continue to use relative entropy, the constrained worst case still entails exponential tilting, but θ becomes a Lagrange multiplier that depends on date t information. The recursion must subtract off $\beta\theta_t$ times the entropy constraint. As demonstrated by Petersen, James and Dupuis (2000) and Hansen et al. (2006), a time invariant parameter θ may be interpreted as a Lagrange multiplier of an intertemporal constraint, in contrast to the specifications advocated by Epstein and Schneider (2003).

The challenge for empirical work becomes one estimating penalization parameters or alternatively the size of constraints on the families of probabilities. These objects replace the incremental risk adjustments.

2.3. Intertemporal complementarity and social externalities

Consider next a specification with intertemporal complementarities. Introduce a habit stock, which we model as evolving according to

$$H_t = (1-\lambda)C_t + \lambda H_{t-1}$$

where λ is a depreciation factor and H_t is a geometric average of current and past consumptions. In building preferences, form an intermediate object that depends on both current consumption and the history of consumption:

$$S_t = \left[\delta(C_t)^{1-\alpha} + (1-\delta)(H_t)^{1-\alpha}\right]^{\frac{1}{1-\alpha}}$$

where $\alpha > 0$ and $0 < \delta < 1$. Construct the continuation value recursively via

$$V_t = \left[(1-\beta)(S_t)^{1-\rho} + \beta\left[\mathsf{R}(V_{t+1}|\mathcal{F}_t)\right]^{1-\rho}\right]^{\frac{1}{1-\rho}}.$$

Alternatively, H_t may be used as a subsistence point in the construction of S_t as in

$$S_t = C_t - \delta H_t.$$

Typically $\mathsf{R}(V_{t+1}|\mathcal{F}_t) = [E(V_{t+1}^{1-\rho}|\mathcal{F}_t)]^{\frac{1}{1-\rho}}$, and this specification is used as a distinct way to separate risk aversion and intertemporal substitution. Intertemporal substitution is now determined by more than just ρ: in particular the preference parameters (δ, α) along with ρ and the depreciation factor λ. The parameter ρ is typically featured as the *risk aversion parameter*.

Preferences of this general type in asset pricing have been used by Novales (1990), Constantinides (1990), Heaton (1995) and others. Novales used it to build an equilibrium model of real interest rates, but used a specification with quadratic adjustment costs in consumption. Instead of using CES specification, Constantinides and Heaton use H_t to shift the subsistence point in the preferences to study the return differences between equity and bonds. It remains an open issue as to how important these various distinctions are in practice.

When the consumer accounts for the effect of the current consumption choice on future values of the habit stock, the habit effects are *internal* to the consumer. Sometimes the habit stock H_t is taken to be *external* and outside the control of the consumer. The habit stock serves as a social reference point. Examples include Abel (1990) and Campbell and Cochrane (1999).

3. Stochastic discount factors

In this section we describe how investor preferences become encoded in asset prices via stochastic discount factors. Our use of stochastic discount factor representations follows Harrison and Kreps (1979) and Hansen and Richard (1987) and many others.

For the time being we focus on one-period pricing and hence one-period stochastic discount factors; but subsequently we will explore multi-period counterparts. Multi-period stochastic discount factors are built by forming products of single period stochastic discount factors.

3.1. One-period pricing

Consider the one-period pricing of elements X_{t+1} in a space of asset payoffs. An asset payoff is a bundled (across states) claim to a consumption numeraire over alternative states of the world that are realized at a future date. Thus payoffs $x_{t+1} \in X_{t+1}$ depend on information available at $t+1$. Mathematically they are depicted as a random variable in the date $t+1$ conditioning information set of investors. The time t price of x_{t+1} is denoted by $\pi_t(x_{t+1})$ and is in the date t information set $\mathcal{F}_t$ of investors.

Hansen and Richard (1987) give restrictions on the set of payoffs and prices for there to exist a representation of the pricing function of the form

$$E(S_{t,t+1}x_{t+1}|\mathcal{F}_t) = \pi_t(x_{t+1}) \tag{2}$$

where $\mathcal{F}_t$ is the current conditioning information set which is common across investors. These restrictions allow investors to use information available at date t to trade in frictionless markets.[3] The positive random variable $S_{t,t+1}$ is a stochastic discount factor used to price assets. It *discounts* asset payoffs differently depending on the realized state in a future time period. Consequently, this *discounting* encompasses both the discounting of known payoffs using a risk-free interest rate and the adjustments for risk. As argued by Harrison and Kreps (1979) and others, the existence of a positive stochastic discount factor follows from the absence of arbitrage opportunities in frictionless markets.

A common and convenient empirical strategy is to link stochastic discount factors to intertemporal marginal rates of substitution. We illustrate this for a two-period economy, but we will deduce formulas for dynamic economies in subsequent presentation.

EXAMPLE 3.1. Suppose that investor j maximizes the utility function

$$E\left[u^j\left(c_t^j, c_{t+1}^j\right)|\mathcal{F}_t\right]$$

by trading financial claims. Let $(\bar{c}_t^j, \bar{c}_{t+1}^j)$ be the optimal consumption choices for this consumer. Consider a perturbation of this consumption bundle in the direction $(\bar{c}_t^j - \mathsf{r}\pi_t(x_{t+1}), \bar{c}_{t+1}^j + \mathsf{r}x_{t+1})$ which is parameterized by the real number r. Notice that this change in consumption is budget neutral for all choices of r. Differentiating with

[3] Hansen and Richard (1987) impose conditional second moment restrictions on payoffs and a specific form of conditional continuity. Other conditional moment and conditional continuity restrictions can also be used to justify this representation.

respect to r, at the optimal choices we have

$$E\left[u_1^j\left(\bar{c}_t^j, \bar{c}_{t+1}^j\right)|\mathcal{F}_t\right]\pi_t(x_{t+1}) = E\left[u_2^j\left(\bar{c}_t^j, \bar{c}_{t+1}^j\right)x_{t+1}|\mathcal{F}_t\right].$$

As a result

$$E\left(M_{t,t+1}^j x_{t+1}|\mathcal{F}_t\right) = \pi_t(x_{t+1}) \tag{3}$$

where the intertemporal marginal rate of substitution:

$$M_{t,t+1}^j \doteq \frac{u_2^j(\bar{c}_t^j, \bar{c}_{t+1}^j)}{E[u_1^j(\bar{c}_t^j, \bar{c}_{t+1}^j)|\mathcal{F}_t]}.$$

This same argument applies to any feasible perturbation and hence (3) is applicable to any payoff as long as the perturbation away from the optimal that we explored is permitted. This gives a link between important economic quantities and asset prices.

Note that

$$E\left[\left(M_{t,t+1}^j - M_{t,t+1}^i\right)x_{t+1}|\mathcal{F}_t\right] = 0$$

for all investors j and i. Therefore any difference in the marginal rates of substitution across agents are orthogonal to the payoff space X_{t+1}.

Suppose now that X_{t+1} includes any bounded function that is measurable with respect to a sigma algebra $\mathcal{G}_{t+1}$ that is contained in $\mathcal{F}_{t+1}$. Then this orthogonality implies:

$$E\left(M_{t,t+1}^j|\mathcal{G}_{t+1}\right) = S_{t,t+1}$$

for all j. The stochastic discount factor is unique if it is restricted to be measurable with respect to $\mathcal{G}_{t+1}$. More generally, any of the intertemporal marginal rates of substitution of the investors can be used as a stochastic discount factor to depict prices. One economically important example of the difference between $\mathcal{G}_{t+1}$ and $\mathcal{F}_{t+1}$ is the case where there are traded claims to aggregate uncertainty but claims to individual risk are not. Therefore there is limited risk-sharing in financial markets in this economy.[4]

Suppose that investors can trade contracts contingent on any information that is available as of date $t+1$. Further suppose that these investors do not face any trading frictions such as transactions costs or short-sale constraints. Under this complete market specification $\mathcal{G}_{t+1} = \mathcal{F}_{t+1}$ and $\mathcal{F}_{t+1}$ includes all individuals' information. In this case

$$M_{t,t+1}^j = S_{t,t+1}$$

and $S_{t,t+1}$ is unique. The marginal rates of substitution are equated across investors.

For pedagogical simplicity we compute shadow prices. That is we presume that consumption is fixed at some determined process. Subsequently, we will have to add specificity to this process, but for the time being we remain a bit agnostic. It can be the outcome of a decentralized production economy, but we place production considerations on the back burner.

[4] See, for example, Constantinides and Duffie (1996).

3.2. CES benchmark

Consider an economy with complete markets and investors with identical preferences of this CES type. In what follows we will use the common preference specification to deduce a formula for the stochastic discount factor. For the recursive utility model with a CES specification, it is convenient to represent pricing in two steps. First we value a contingent claim to next period's continuation value. We then change units from continuation values to consumption by using the next-period marginal utility for consumption. In all cases, marginal utilities are evaluated at aggregate consumption. The CES specification makes these calculations easy and direct.

Because the CES recursion is homogeneous of degree one in its arguments, we can use Euler's Theorem to write

$$V_t = (MC_t)C_t + E\big[(MV_{t+1})V_{t+1}|\mathcal{F}_t\big]. \tag{4}$$

Claims to future continuation values V_{t+1} can be taken as substitutes for claims to future consumption processes. When current consumption be the numeraire, equilibrium wealth is given by $W_t \equiv V_t/MC_t$. Divide (4) by MC_t to obtain a recursive expression for wealth:

$$W_t = C_t + E[S_{t,t+1}W_{t+1}|\mathcal{F}_t].$$

The marginal utility of consumption is

$$MC_t = (1-\beta)(C_t)^{-\rho}(V_t)^{\rho},$$

and the marginal utility of next-period continuation value is

$$MV_{t+1} = \beta(V_{t+1})^{-\gamma}\big[\mathsf{R}(V_{t+1}|\mathcal{F}_t)\big]^{\gamma-\rho}(V_t)^{\rho}. \tag{5}$$

Forming the intertemporal marginal rate of substitution gives

$$S_{t,t+1} = \beta\left(\frac{C_{t+1}}{C_t}\right)^{-\rho}\left[\frac{V_{t+1}}{\mathsf{R}(V_{t+1}|\mathcal{F}_t)}\right]^{\rho-\gamma}.$$

When we incorporate kinks in preferences as in setting suggested by Routledge and Zin (2003), the marginal utility of next-period continuation value is

$$MV_{t+1} = \beta(V_{t+1})^{-\gamma}\big[\mathsf{R}(V_{t+1}|\mathcal{F}_t)\big]^{\gamma-\rho}(V_t)^{\rho} \times\left[\frac{1+\alpha\mathbf{1}_{\{V_{t+1}\leqslant\delta\mathsf{R}(V_{t+1}|\mathcal{F}_t)\}}}{1+\delta^{1-\gamma}\alpha E(\mathbf{1}_{\{V_{t+1}\leqslant\delta\mathsf{R}(V_{t+1}|\mathcal{F}_t)\}}|\mathcal{F}_t)}\right].$$

Combining these terms, the one-period intertemporal marginal rate of substitution is

$$S_{t,t+1} = \beta\left(\frac{C_{t+1}}{C_t}\right)^{-\rho}\left[\frac{V_{t+1}}{\mathsf{R}(V_{t+1}|\mathcal{F}_t)}\right]^{\rho-\gamma} \times\left[\frac{1+\alpha\mathbf{1}_{\{V_{t+1}\leqslant\delta\mathsf{R}(V_{t+1}|\mathcal{F}_t)\}}}{1+\delta^{1-\gamma}\alpha E(\mathbf{1}_{\{V_{t+1}\leqslant\delta\mathsf{R}(V_{t+1}|\mathcal{F}_t)\}}|\mathcal{F}_t)}\right].$$

The stochastic discount factor depends directly on current consumption, and indirectly on future consumption through the continuation value.

We now consider some special cases of the CES version of the Kreps–Porteus model:

EXAMPLE 3.2. Let $\rho = \gamma$ and $\alpha = 0$. Then the contribution to the continuation value drops out from the stochastic discount factor. This is the model of Lucas (1978) and Breeden (1979).

EXAMPLE 3.3. Consider the special case with $\rho = 1$ and $\alpha = 0$, but allow γ to be distinct from one. Then the counterpart to the term $\left(\frac{V_{t+1}}{\mathcal{R}_t(V_{t+1}|\mathcal{F}_t)}\right)^{\rho-\gamma}$ entering the stochastic discount factor is

$$\frac{(V_{t+1})^{1-\gamma}}{E[(V_{t+1})^{1-\gamma}|\mathcal{F}_t]}.$$

Notice that this term has conditional expectation equal to unity.

EXAMPLE 3.4. Consider the special case in which $\gamma = 1$ and $\alpha = 0$, but allow ρ to be distinct from one. In this case the counterpart to the term $\left(\frac{V_{t+1}}{\mathcal{R}_t(V_{t+1}|\mathcal{F}_t)}\right)^{\rho-\gamma}$ entering the stochastic discount factor is

$$\left[\frac{V_{t+1}}{\exp E(\log V_{t+1}|\mathcal{F}_t)}\right]^{\rho-1}.$$

The logarithm of this term has expectation zero.

4. Empirical observations from asset returns

Time series observations of asset returns and consumption are needed to identify the parameters governing the preferences of consumers. The stochastic discount factor developed in Section 3 and its implications for security prices impose a set of joint restrictions on asset prices and consumption. Before analyzing these restrictions, we first display some important empirical regularities from asset markets alone. Besides standard sample statistics for asset returns we also examine some standard decompositions of prices. These are based on a log-linear approximation and the present-value relationship.

This decomposition was proposed by Campbell and Shiller (1988a, 1988b) and Cochrane (1992). The methods have been used extensively in the finance literature to summarize statistical evidence about dividend–price ratios, dividend growth and returns. We develop these methods and show their link to related work in the macroeconomics literature by Hansen, Roberds and Sargent (1991). We then apply these decompositions to an important set of test assets.

4.1. Log linear approximation and present values

The price of a security at time t is given by P_t. The return to this security from time t to time $t+1$ is determined by the cash flow received at time $t+1$, denoted D_{t+1} and the price of the security at time $t+1$, denoted P_{t+1}. The return is given by

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t} = \left(\frac{D_{t+1}}{D_t}\right)\left(\frac{1 + P_{t+1}/D_{t+1}}{P_t/D_t}\right). \tag{6}$$

The cash flow, D_{t+1} is the dividend in the case of stocks or a coupon in the case of bonds. Although many individual companies do not pay dividends, our empirical analysis is based on the analysis of portfolios of stocks and these dividends will be positive.

This allows us to take logarithms of (6). Using lower case letters to denote logarithms of each variable we have

$$r_{t+1} = (d_{t+1} - d_t) - (p_t - d_t) + \log\bigl[1 + \exp(p_{t+1} - d_{t+1})\bigr]. \tag{7}$$

We view this as a difference equation for the logarithm of the price–dividend ratio with *forcing processes* given by the returns and dividend growth rate. The use of returns as a forcing process allow us to deduce some statistical restrictions. The valuation models of Section 3 determine both the prices and the returns endogenously.

To make (7) a linear difference equation, consider the approximation

$$\log\bigl[1 + \exp(p_{t+1} - d_{t+1})\bigr] \approx \log\bigl[1 + \exp(\mu_{p-d})\bigr] + \kappa(p_{t+1} - d_{t+1} - \mu_{p-d}) \tag{8}$$

where

$$\kappa = \frac{\exp(\mu_{p-d})}{1 + \exp(\mu_{p-d})} < 1$$

and μ_{p-d} is a constant centering point for the linearization. This point is usually taken to be the mean of the logarithm of the price–dividend ratio which will be different for alternative cash flows because of differences in cash flows and discount rates.

Substitute approximation (8) into the difference equation (7) and rearrange terms:

$$p_t - d_t = (d_{t+1} - d_t) - r_{t+1} + \kappa(p_{t+1} - d_{t+1}) + \mathsf{c} \tag{9}$$

where

$$\mathsf{c} = \log\bigl[1 + \exp(\mu_{p-d})\bigr] - \kappa\mu_{p-d}.$$

For the remainder of this section, we will ignore the approximation error and treat (9) as the difference equation of interest.

Solving (9) forward gives

$$p_t - d_t = \sum_{j=0}^{\infty} (\kappa)^j [d_{t+j+1} - d_{t+j} - r_{t+j+1}] + \frac{\mathsf{c}}{1-\kappa}.$$

Notice that the constant term in this solution satisfies the approximation

$$\frac{\mathsf{c}}{1-\kappa} \approx \mu_{p-d}.$$

4.1.1. Moving-average models

The implications of the linear difference equation for returns will be examined using simple linear time series models. We therefore assume that there is a first-order Markov process for a state vector x_t where the dynamics are given by

$$x_{t+1} = Ax_t + Bw_{t+1} \tag{10}$$

where $\{w_{t+1}\colon t = \ldots, 0, 1, \ldots\}$ is a sequence of *iid* normally distributed random vectors with $E[w_{t+1}] = 0$ and $E(w_{t+1}w'_{t+1}) = I$. The matrix A is assumed to have eigenvalues with absolute values less than 1. This assumption implies a stochastic steady state for x_t where x_t is a moving-average of current and past shocks:

$$x_t = \sum_{j=0}^{\infty} A^j B w_{t-j} = \sum_{j=0}^{\infty} A^j B L^j w_t \equiv \mathcal{A}(L) w_t$$

where L denotes the "lag" operator.

Dividends, returns and prices are linked to the state vector x_t via

$$\begin{aligned} d_{t+1} - d_t &= \mu_d + G_d x_t + H_d w_{t+1}, \\ r_{t+1} &= \mu_r + G_r x_t + H_r w_{t+1}, \\ p_t - d_t &= \mu_{p-d} + G_{p-d} x_t. \end{aligned}$$

The present-value model implies restrictions on this representation, which we now explore. We will derive these restrictions in two ways. Substitute these depictions into (9) and obtain:

$$\begin{aligned} G_{p-d} x_t &= (G_d - G_r + \kappa G_{p-d} A) x_t, \\ 0 &= (H_d - H_r + \kappa G_{p-d} B) w_{t+1}. \end{aligned}$$

Since these restrictions must hold for all realized values of x_t and w_{t+1}, these two equations restrict directly the representation for dividends, returns and the price–dividend ratio.

To obtain an alternative perspective on these restrictions, we use the implied moving-average representations. In stochastic steady state, dividends and returns satisfy

$$\begin{aligned} d_{t+1} - d_t &= \delta(L) w_{t+1} + \mu_d, \\ r_{t+1} &= \rho(L) w_{t+1} + \mu_r, \\ p_t - d_t &= \pi(L) w_t + \mu_{p-d}, \end{aligned}$$

where

$$\delta(z) = \sum_{j=0}^{\infty} \delta_j z^j, \quad \sum_{j=0}^{\infty} |\delta_j|^2 < \infty,$$

$$\rho(z) = \sum_{j=0}^{\infty} \rho_j z^j, \quad \sum_{j=0}^{\infty} |\rho_j|^2 < \infty,$$

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j, \quad \sum_{j=0}^{\infty} |\pi_j|^2 < \infty.$$

The variable z is introduced so that we may view $\delta(z)$, $\rho(z)$, $\pi(z)$ as power series. They are sometimes referred to as the z-transforms of the moving-average coefficients. The coefficients of the power series are the moving-average coefficients. The power series converge at least on the domain $|z| < 1$.

In this case, the coefficients of the power series $\delta(z)$ and $\rho(z)$ are given by

$$\delta_0 = H_d, \qquad \rho_0 = H_r,$$

$$\delta_j = G_d A^{j-1} B, \quad \rho_j = G_r A^{j-1} B.$$

Hence

$$\delta(z) = H_d + zG_d(I - zA)^{-1}B,$$

$$\rho(z) = H_r + zG_r(I - zA)^{-1}B.$$

Difference equation (9) implies that

$$z\pi(z) = \delta(z) - \rho(z) + \kappa\pi(z). \tag{11}$$

This is an equation that restricts the moving average coefficients. We may evaluate these functions at $z = \kappa$:

$$\kappa\pi(\kappa) = \delta(\kappa) - \rho(\kappa) + \kappa\pi(\kappa).$$

This implies that

$$\delta(\kappa) = \rho(\kappa). \tag{12}$$

Using the power series representation of ρ and δ, this implies that the discounted (by κ) impulse responses for returns and cash flow growth rates must be equal. This is the present-value–budget-balance restriction of Hansen, Roberds and Sargent (1991). This restriction is necessary in order that the future shocks to cash flow growth rates and to returns net out so that the price–dividend ratio only depends on current and past shocks.

Under the Markov representation of the state variable x_t, the restriction

$$\rho(\kappa) = \delta(\kappa)$$

becomes

$$H_r + \kappa G_r (I - \kappa A)^{-1} B = H_d + \kappa G_d (I - \kappa A)^{-1} B.$$

The moving average representation for the price–dividend ratio is obtained by solving Equation (11) for π:

$$\pi(z) = \frac{\delta(z) - \rho(z)}{z - \kappa}. \tag{13}$$

Because of the denominator term, the right-hand side looks like it explodes at $z = \kappa$. This is not the case, however. The numerator is also zero at $z = \kappa$. After dividing out the common zero at κ, π will have a well-defined power series for $|z| < 1$, and formula (13) for $\pi(z)$ is a valid formula for the z-transform of the moving-average coefficients. Performing this division is consistent with the formula

$$G_{p-d} = (G_d - G_r)(I - \kappa A)^{-1}$$

used in representing the price–dividend ratio.

This "solution" is a bit unusual. It takes returns and dividend growth as given and solves for the price–dividend ratio. A structural asset pricing model does in fact have different primitives. Even when cash flows are given exogenously, returns and price–dividend ratios are both determined endogenously. The rationale for "solving" the model in this manner is instead a way to allow for prices or returns to reveal additional information used by investors to forecast future cash flows. It is a restriction imposed on a moving-average representation of the shocks that are pertinent to the investors' decision-making.

4.1.2. Decompositions

This solution for π is often used to motivate empirical decompositions of prices and measurement of return risk.

1. *Return decomposition*. The risk in returns from time t to time $t + 1$ is captured by the term $\rho_0 w_{t+1}$. Since $\rho(\kappa) = \delta(\kappa)$,

$$\rho_0 = \delta(\kappa) - \sum_{j=1}^{\infty} \kappa^j \rho_j.$$

Hence one period exposure to risk has both a discounted cash flow component and a component due to return predictability. When return predictability is not very strong, the discounted impact of shocks on future dividends is the most important source of risk. In addition if κ is close to one, $\delta(\kappa)$ measures the accumulated impact of current shocks on dividends far into the future. This measure of long-run risk is featured in the work of Bansal, Dittmar and Lundblad (2005) and Hansen, Heaton and Li (2005).

2. *Price–dividend decomposition*. Using $\rho(\kappa) = \delta(\kappa)$, express π as

$$\pi(z) = \left[\frac{\delta(z) - \delta(\kappa)}{z - \kappa}\right] - \left[\frac{\rho(z) - \rho(\kappa)}{z - \kappa}\right].$$

The first term is the discounted expected future cash flow growth and the second is the discounted expected future returns both net of constants. This decomposition is used to measure the importance of discounted cash flows in accounting for variation in the price–dividend ratio. This decomposition was originally proposed by Campbell and Shiller (1988a, 1988b).

4.1.3. Identifying shocks

For the restriction on the joint dynamics of returns, dividends and prices to be testable, we must be able to identify shocks. Vector autoregressive (VAR) methods are commonly used in conjunction with other restrictions to identify shocks. Hansen, Roberds and Sargent (1991) show that there is a tension, however, between the use of VAR methods to identify shocks and the present-value-budget-balance implications that are imposed in the log-linear model.

Let y_t be a vector of observables with moving average representation:

$$y_{t+1} = \mathcal{B}(L)w_{t+1} + \mu_y.$$

To construct w_{t+1} from $y_{t+1}, y_t, \ldots$ requires that $\mathcal{B}(z)$ be of full rank for $|z| < 1$. In vector autoregressive applications, it is typically assumed that y and w have the same number of entries. In this case $\mathcal{B}(z)$ must be nonsingular for $z < 1$, and, in particular, $\mathcal{B}(\kappa)$ must be nonsingular. If y_{t+1} contains $d_{t+1} - d_t$ and r_{t+1} as the first two entries, then $\delta(\kappa) = \rho(\kappa)$ implies that

$$[\,1 \quad -1 \quad 0\,]\mathcal{B}(\kappa) = 0$$

which violates the restriction that $\mathcal{B}(z)$ be nonsingular. Returns do not contain enough information to reveal shocks along with dividend growth. This is the dividend-return counterpart to a claim established in Hansen, Roberds and Sargent (1991), and it gives a warning against using VAR methods in conjunction with dividends and returns alone.

Let y_{t+1} include cash flow growth rates $d_{t+1} - d_t$ and the price–dividend ratio $p_{t+1} - d_{t+1}$. Given the implied moving-average representations from a state-space model or a VAR form:

$$p_t - d_t = \pi(L)w_t + \mu_{p-d},$$
$$d_{t+1} - d_t = \delta(L)w_{t+1} + \mu_d.$$

In this case construct the moving-average representation for the approximate return via

$$r_{t+1} = \rho(L)w_{t+1} + \mu_r$$

where $\rho(z) = \delta(z) + (\kappa - z)\pi(z)$. This necessarily satisfies the present-value restriction (12). Thus we sidestep the informational inconsistency mentioned previously by using prices to reveal shock components instead of returns.

4.2. Test assets

To illustrate the construction of these returns we use the prices, returns and dividends constructed from six portfolios. The portfolios returns and dividends are constructed as in Hansen, Heaton and Li (2005).

The first portfolio is a market portfolio of stocks traded on the NYSE and NASDAQ. The other portfolios are constructed by sorting stocks on the basis of book value relative to market value of equity as in Fama and French (1992). Five portfolios with equal numbers of stocks in each portfolio are constructed from the entire universe of stocks. Dividends are then constructed from the return series for each portfolio with and without dividends. This construction is done on a quarterly basis from 1947 to 2005. Because of the pronounced seasonality in dividends, dividends are smoothed over a year. Details of the data construction can be found in Hansen, Heaton and Li (2005).

Table 1 reports summary statistics for the five book-to-market portfolios (portfolios "1" through "5"). Notice that portfolio 1 has the lowest average book-to-market value (B/M) and the highest average price–dividend ratio (P/D) and the lowest average return. Moving from portfolio 1 to portfolio 5, the average book-to-market value increases, the average price–dividend ratio declines and the average return increases. As we will see in Section 6, differences in the average returns are not explained by exposure to contemporaneous covariance with consumption.

Table 1
Properties of portfolios sorted by book-to-market

	Portfolio					
	1	2	3	4	5	Market
One-period exp. return (%)	6.79	7.08	9.54	9.94	11.92	7.55
Long-run return (%)	8.56	8.16	10.72	10.84	13.01	8.77
Avg. B/M	0.32	0.61	0.83	1.10	1.80	0.65
Avg. P/D	51.38	34.13	29.02	26.44	27.68	32.39

Notes. Data are quarterly from 1947 Q1 to 2005 Q4 for returns and annual from 1947 to 2005 for B/M ratios. Returns are converted to real units using the implicit price deflator for nondurable and services consumption. Average returns are converted to annual units using the natural logarithm of quarterly gross returns multiplied by 4. "One-period exp. return," we report the predicted quarterly gross returns to holding each portfolio in annual units. The expected returns are constructed using a separate VAR for each portfolio with inputs $(c_t - c_{t-1}, e_t - c_t, r_t)$ where r_t is the logarithm of the gross return of the portfolio. We imposed the restriction that consumption and earnings are not Granger caused by the returns. One-period expected gross returns are calculated conditional on being at the mean of the state variable implied by the VAR. "Long-run return" reports the limiting value of the logarithm of the expected long-horizon return from the VAR divided by the horizon. "Avg. B/M" for each portfolio is the average portfolio book-to-market over the period computed from COMPUSTAT. "Avg. P/D" gives the average price–dividend for each portfolio where dividends are in annual units.

4.2.1. Vector autoregression

We first consider a statistical decomposition of the price–dividend ratio for each portfolio using vector autoregressions. To do this let

$$y_t \equiv \begin{bmatrix} d_t - d_{t-1} \\ p_t - d_t \end{bmatrix}.$$

We fit a VAR of the form

$$y_t = A_0 + A_1 y_{t-1} + \cdots + A_l y_{t-l} + B w_t$$

where the two-dimensional shock vector w_t has mean zero and covariance matrix I. Further A_0 is two-dimensional, the matrices A_j, $j = 1, 2, \ldots, l$, and B are two by two. We further impose the normalization that B is lower triangular so that the second shock (the second element of w_t) does not impact dividend growth contemporaneously.

This VAR implies linear dynamics for the Markov process x_t. To see this, let

$$\mu \equiv E(y_t) = (I - A_1 - \cdots A_l)^{-1} A_0$$

and

$$y_t^* \equiv y_t - \mu.$$

Then x_t is given by

$$x_t \equiv \begin{bmatrix} y_t^* \\ y_{t-1}^* \\ \vdots \\ y_{t-l}^* \end{bmatrix}, \quad G \equiv \begin{bmatrix} A_1 A_2 \cdots A_l \\ I \;\; 0 \cdots 0 \\ \vdots \;\; \ddots \;\; \vdots \;\; \vdots \\ 0 \cdots I \;\; 0 \end{bmatrix} \quad \text{and} \quad H \equiv \begin{bmatrix} B \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

For each portfolio we estimate a VAR with $l = 5$ and consider the properties of portfolio cash flows and prices using estimated impulse response functions.

Figure 1 reports the impulse response functions for the market. The top panel of the figure reports the response of the level of log dividends to the two shocks. The first shock has an immediate effect on dividends and then the response builds going forward. The second shock has a very small effect on dividends. The second panel of the figure reports the response of the log price–dividend ratio to the shocks. Notice that the first shock has a very little effect on the price–dividend ratio, while the second shock increases the price–dividend ratio and the impact persists for many periods. The pattern of responses indicates that the two shocks can be labeled as a dividend shocks and a separate price–dividend shock. Shocks to the price–dividend ratio are long-lasting and have little ability to forecast future dividends. This reflects the well-known inability of the price–dividend ratio at the aggregate level to forecast future dividends.

The bottom panel of Figure 1 reports the implied response of returns to the two shocks. To better understand the effects of the shocks, the results are reported for the *cumulative* impact of the shocks on returns. Notice that the dividend shock (shock 1)

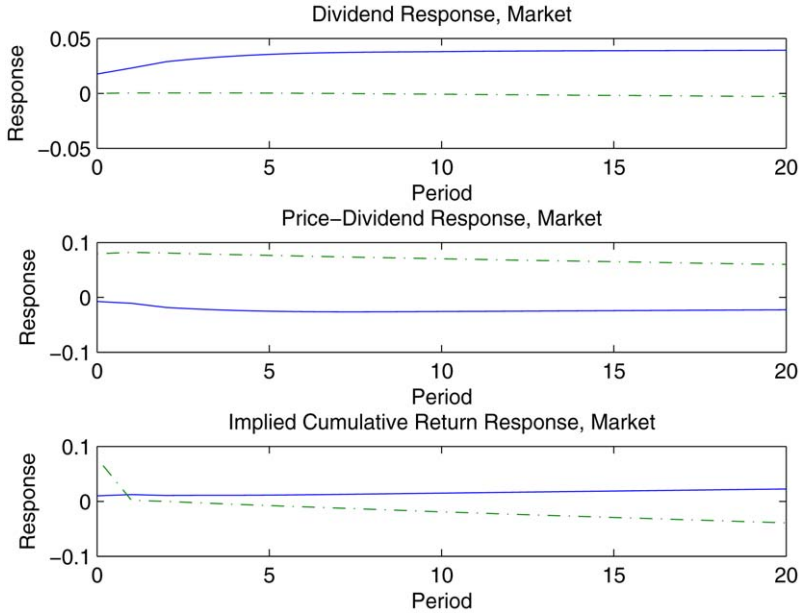

Figure 1. Impulse response functions for the market portfolio. Top panel: response of log dividends to shocks. Middle panel: response of the log price–dividend ratio to shocks. Bottom panel: response of returns to shocks. — depicts impulse responses to the first shock. – · – depicts impulse responses to the second shock.

has little effect on returns while the price–dividend shock has an initial positive impact on returns followed by a slowly building negative impact on returns in the future. For the market portfolio, variation in the price–dividend ratio has some predictive ability for future returns, while variation in dividends that have no effect on prices, has little ability to forecast future returns.

These results are interpreted by Campbell and Shiller (1988a, 1988b) and others as implying that variation in future returns is the most important factor explaining variation in the price–dividend ratio. Further this variation is empirically independent of variation in future dividends. This implies that for this aggregate portfolio variation in the price–dividend ratio must be driven by required returns. This has potentially important implications for the stochastic discount factor of Section 3.

The corresponding impulse response functions for portfolios 1 and 5 are reported in Figures 2 and 3, respectively. Notice that for these portfolios the labeling of the two shocks as dividend and return shocks is not clear. For example, shocks to dividends now have an ability to forecast future returns. As portfolio returns and dividends are disaggregated, the predictability of dividends rises. This fact is emphasized in the work of Vuolteenaho (2002).

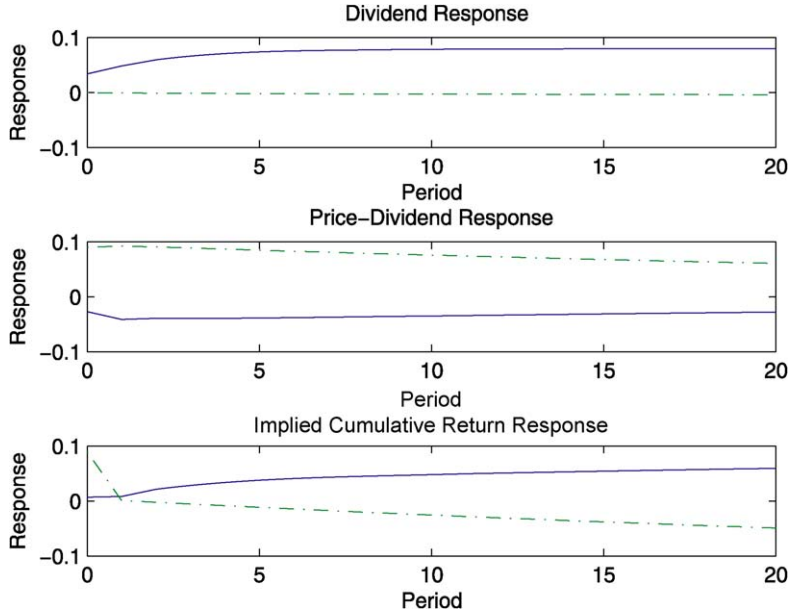

Figure 2. Impulse response functions for the portfolio 1. Top panel: response of log dividends to shocks. Middle panel: response of the log price–dividend ratio to shocks. Bottom panel: response of returns to shocks. — depicts impulse responses to the first shock. – · – depicts impulse responses to the second shock.

5. Intertemporal substitution and pricing

To understand how investor preference parameters and the stochastic environment influence asset prices, we explore further the solution of the CES version of the Kreps–Porteus model for fixed, prespecified consumption process as in a Lucas-style *endowment economy*. We derive some approximation results where we approximate around a unitary intertemporal substitution parameter $\rho = 1$ for an arbitrary value of $\gamma > 0$. Thus we feature the role of this parameter in our characterizations. As in Restoy and Weil (1998) consumption dynamics plays a central role in these characterizations. For some specifications of consumption dynamics we obtain a structural model of the type analyzed in Section 4.

Our expansion in ρ follows in part the work of Kogan and Uppal (2001).[5] The economy we study is different from that of Kogan and Uppal (2001), but they suggest that extensions such as those developed here would be fruitful. By approximating around ρ, we are approximating around a stochastic economy with a constant consumption wealth ratio. As we will see, the $\rho = 1$ limit economy leads to other less dramatic

[5] Our ρ derivatives will be heuristic in the sense that we will not provide a rigorous development of their approximation properties.

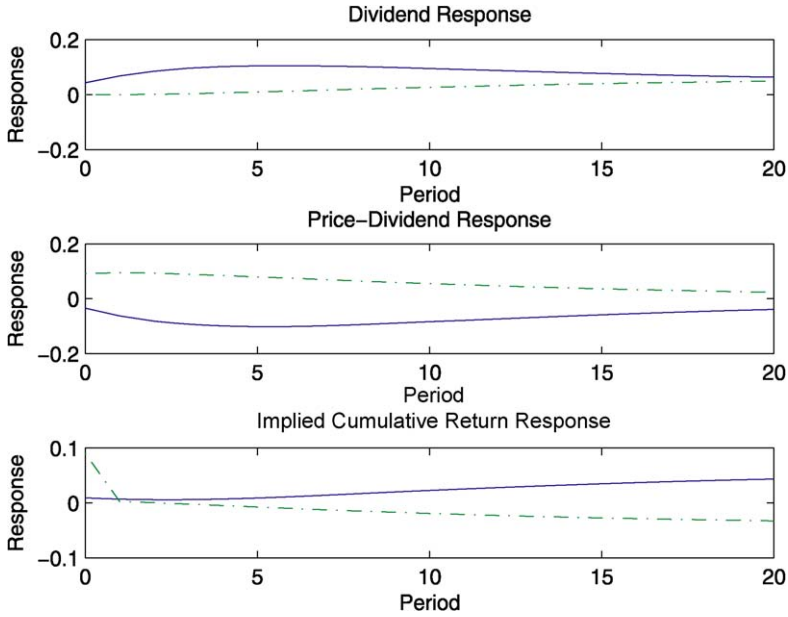

Figure 3. Impulse response functions for the portfolio 5. Top panel: response of log dividends to shocks. Middle panel: response of the log price–dividend ratio to shocks. Bottom panel: response of returns to shocks. — depicts impulse responses to the first shock. – · – depicts impulse responses to the second shock.

simplifications that we exploit in characterizing asset prices and risk premia. The simplifications carry over the ρ derivatives that we calculate for asset prices and returns. While Campbell and Viceira (2002, Chapter 5) show the close connection between approximation around the utility parameter $\rho = 1$ and approximation around a constant consumption-wealth ratio for portfolio problems, there are some interesting differences in our application. Moreover, $\rho = 1$ is inconveniently ruled out in the parameterization of recursive utility considered by Restoy and Weil (1998) and others because of their use of the return-based Euler equation.

We consider first a family of discrete-time economies with log-linear consumption dynamics indexed by ρ. When we introduce stochastic volatility in consumption, we find it more convenient to explore a family of economies specified in continuous time. We illustrate these economies using some parameter values extracted from existing research.

5.1. Discrete time

The initial step in our calculation is the first-order expansion of the continuation values in the parameter ρ. Let v_t^ρ denote the logarithm of the continuation value for intertemporal substitution parameter ρ, and let c_t denote the logarithm of consumption. We rewrite

the CES recursion as

$$v_t^\rho = \frac{1}{1-\rho}\log\bigl[(1-\beta)\exp\bigl[(1-\rho)c_t\bigr] + \beta\exp\bigl[(1-\rho)\mathcal{Q}_t\bigl(v_{t+1}^\rho\bigr)\bigr]\bigr], \tag{14}$$

where $\mathcal{Q}_t$ is

$$\mathcal{Q}_t(v_{t+1}) = \frac{1}{1-\gamma}\log E\bigl(\exp\bigl[(1-\gamma)v_{t+1}\bigr]|\mathcal{F}_t\bigr).$$

When $\rho = 1$ this recursion simplifies to

$$v_t^1 = (1-\beta)c_t + \beta\mathcal{Q}_t\bigl(v_{t+1}^1\bigr). \tag{15}$$

5.1.1. Continuation values

We compute the first-order expansion

$$v_t^\rho \approx v_t^1 + (\rho-1)Dv_t^1$$

where v_t^1 is the continuation value for the case in which $\rho = 1$ and the notation D denotes the differentiation with respect to ρ. We construct an approximate recursion for Dv_t^1 by expanding the logarithm and exponential functions in (14) and including up to second-order terms in c_t and $\mathcal{Q}_t$. The approximate recursion is:

$$v_t^\rho \approx (1-\beta)c_t + \beta\mathcal{Q}_t\bigl(v_{t+1}^\rho\bigr) + \beta(1-\rho)(1-\beta)\frac{[\mathcal{Q}_t(v_{t+1}^\rho) - c_t]^2}{2}. \tag{16}$$

As is evident from (15), this approximation is exact when $\rho = 1$.

Our aim is to construct an exact recursion for the derivative of v_t with respect to ρ. One way to do this is to differentiate directly (14). It is simpler to differentiate the approximate recursion (16) for the logarithm of the continuation value v_t with respect to ρ. This is valid because the approximation error in the recursion does not alter the derivative with respect to ρ. Performing either calculation gives

$$\begin{aligned} Dv_t^1 &= -\beta(1-\beta)\frac{[\mathcal{Q}_t(v_{t+1}^1) - c_t]^2}{2} + \beta E^*\bigl(Dv_{t+1}^1|\mathcal{F}_t\bigr) \\ &= -\frac{(1-\beta)(v_t^1 - c_t)^2}{2\beta} + \beta E^*\bigl(Dv_{t+1}^1|\mathcal{F}_t\bigr) \end{aligned} \tag{17}$$

where E^* is the distorted expectation operator associated with a Radon–Nikodym *derivative*

$$M_{t,t+1} = \frac{\exp[(1-\gamma)v_{t+1}^1]}{E(\exp[(1-\gamma)v_{t+1}^1]|\mathcal{F}_t)}. \tag{18}$$

The Radon–Nikodym derivative is a measure-theoretic notion of a derivative. Since $M_{t,t+1}$ is a positive random variable with conditional expectation one, it induces a distorted probability by scaling random variables. For instance, the distorted expectation

of a random variable is

$$E^*(z_{t+1}|\mathcal{F}_t) = E(M_{t,t+1}z_{t+1}|\mathcal{F}_t).$$

Solving recursion (17) forward gives the derivative Dv_t^1. This derivative is necessarily negative. By using the distorted expectation operator E^* to depict the recursion for Dv_t^1, the recursion has a familiar form that is convenient for computing solutions.

To implement this approach we must compute v_t^1 and the distorted conditional expectation E^*, which will allow us to the solve (17) for Dv_t^1. Later we give some examples when this is straightforward.

5.1.2. Wealth expansion

When ρ is different from one, the wealth–consumption ratio is not constant. Write

$$W_t = \frac{V_t^\rho}{(1-\beta)(C_t)^{-\rho}(V_t^\rho)^\rho} = \frac{(C_t)^\rho (V_t^\rho)^{1-\rho}}{1-\beta}.$$

A first-order expansion of the continuation value implies a second-order expansion of the wealth–consumption ratio. This can be seen by taking logarithms and substituting in the first-order approximation for the continuation value:

$$\begin{aligned}\log W_t - \log C_t &= -\log(1-\beta) + (1-\rho)\left[v_t^1 - c_t + (\rho-1)Dv_t^1\right] \\ &= -\log(1-\beta) - (\rho-1)\left(v_t^1 - c_t\right) - (\rho-1)^2 Dv_t^1. \end{aligned} \tag{19}$$

The first-order term of (19) compares the logarithm of the continuation value for $\rho = 1$ with the logarithm of consumption. The continuation value is forward looking and time varying. Thus when future looks good relative to the present, the term $v_t^1 - c_t$ can be expected to be positive. When the intertemporal elasticity parameter ρ exceeds one, the first-order term implies that a promising future relative to the present has an adverse impact on equilibrium wealth and conversely when ρ is less than one. As we will see, the term v_t^1 is very similar to (but not identical to) the term typically used when taking log-linear approximations.[6]

By construction, the second-order term adjusts the wealth–consumption ratio in a manner that is symmetric about $\rho = 1$, and it is always positive.

5.1.3. Stochastic discount factor expansion

Consider next the first-order expansion of the logarithm of the stochastic discount factor:

$$s_{t+1,t}^\rho \approx s_{t+1,t}^1 + (\rho-1)Ds_{t+1,t}^1.$$

[6] In log-linear approximation the discount rate in this approximation is linked to the mean of the wealth consumption ratio. In the ρ expansion, the subjective rate of discount is used instead.

Recall that the log discount factor is given by

$$s^{\rho}_{t+1,t} = \log\beta - \rho(c_{t+1} - c_t) + (\rho - \gamma)\big[v^{\rho}_{t+1} - \mathcal{Q}_t\big(v^{\rho}_{t+1}\big)\big].$$

Differentiating with respect to ρ gives

$$\begin{aligned} Ds^1_{t+1,t} &= -(c_{t+1} - c_t) + \big[v^1_{t+1} - \mathcal{Q}_t\big(v^1_{t+1}\big)\big] \\ &\quad + (1-\gamma)\big[Dv^1_{t+1} - E^*\big(Dv^1_{t+1}|\mathcal{F}_t\big)\big]. \end{aligned} \tag{20}$$

Thus we obtain the approximation:

$$\begin{aligned} s^{\rho}_{t,t+1} &\approx s^1_{t,t+1} + (\rho - 1)Ds_{t+1,t} \\ &= \log\beta - \rho(c_{t+1} - c_t) + (\rho - \gamma)\big[v^1_{t+1} - \mathcal{Q}_t(v^1_{t+1})\big] \\ &\quad + (1-\gamma)(\rho - 1)\big[Dv^1_{t+1} - E^*(Dv^1_{t+1}|\mathcal{F}_t)\big]. \end{aligned}$$

This shows how changes in ρ alter one period risk prices. For instance consider approximating one period prices of contingent claim z_{t+1} to consumption:

$$\begin{aligned} E\big[\exp\big(s^{\rho}_{t,t+1}\big)z_{t+1}|\mathcal{F}_t\big] &= E\big[\exp\big(s^1_{t,t+1}\big)z_{t+1}|\mathcal{F}_t\big] \\ &\quad + (\rho - 1)E\big[\exp\big(s^1_{t,t+1}\big)Ds_{t,t+1}z_{t+1}|\mathcal{F}_t\big]. \end{aligned}$$

We will explore the ramifications for local risk prices subsequently when we consider a continuous time counterpart to these expansions. This will provide us with formulas for how ρ alters risk premia.

5.1.4. Log-linear dynamics

To show how the previous formulas can be applied, consider the following evolution for consumption in the log linear Markov economy:

$$\begin{aligned} x_{t+1} &= Ax_t + Bw_{t+1}, \\ c_{t+1} - c_t &= \mu_c + G'x_t + H'w_{t+1}, \end{aligned}$$

where $\{w_{t+1}\colon\ t = 0, 1, \ldots\}$ is an iid sequence of standard normally distributed random vectors. Recall that for $\rho = 1$, the continuation value must solve

$$v^1_t = (1-\beta)c_t + \beta\mathcal{Q}_t\big(v^1_{t+1}\big).$$

Conjecture a continuation value of the form

$$v^1_t = U_v \cdot x_t + \mu_v + c_t.$$

Given this guess and the assumed normality,

$$\mathcal{Q}_t(v^1_{t+1}) = U_v'Ax_t + \mu_c + \mu_v + G'x_t + c_t + \frac{1-\gamma}{2}\big|U_v'B + H'\big|^2.$$

Thus

$$U_v = \beta A' U_v + \beta G$$

and

$$\mu_v = \beta\left[\mu_c + \mu_v + \frac{1-\gamma}{2}\left|U_v' B + H'\right|^2\right].$$

Solving for U_v and μ_v,

$$\begin{aligned} U_v &\doteq \beta(I - \beta A')^{-1} G, \\ \mu_v &\doteq \frac{\beta}{1-\beta}\left[\mu_c + \frac{(1-\gamma)}{2}\left|H' + \beta G'(I - A\beta)^{-1} B\right|^2\right]. \end{aligned} \tag{21}$$

For $\rho = 1$ the formulas for the continuation value have simple interpretations. The formula for U_v is also the solution to the problem of forecasting the discounted value of future consumption growth:

$$\begin{aligned} U_v \cdot x_t &= \sum_{j=1}^{\infty} \beta^j E(c_{t+j} - c_{t+j-1} - \mu_c | x_t) \\ &= (1-\beta)\sum_{j=1}^{\infty} \beta^j E(c_{t+j}|\mathcal{F}_t) - \beta c_t - \left(\frac{\beta}{1-\beta}\right)\mu_c. \end{aligned}$$

Therefore,

$$v_t^1 = (1-\beta)\sum_{j=0}^{\infty} \beta^j E(c_{t+j}|\mathcal{F}_t) + \frac{\beta(1-\gamma)}{2(1-\beta)}\left|H' + \beta G'(I - A\beta)^{-1} B\right|^2.$$

The log of the continuation value is a geometric weighted average of logarithms of current and future consumption using the subjective discount factor in the weighting. In addition there is a constant risk adjustment. When consumption growth rates are predictable, they will induce movement in the wealth–consumption ratio as reflected in formula (19). The coefficient on the first-order term in $\rho - 1$ compares the expected discounted average of future log consumption to current log consumption. If this geometric average future consumption is higher than current consumption and ρ exceeds one, the optimistic future induces a negative movement in the wealth–consumption ratio. Conversely a relatively optimistic future induces a positive movement in the wealth–consumption ratio when ρ is less than one.

The constant risk correction term

$$\frac{\beta(1-\gamma)}{2(1-\beta)}\left|H' + \beta G'(I - A\beta)^{-1} B\right|^2$$

entering the continuation value is negative for large values of γ. Consequently, this adjustment enhances the wealth consumption ratio when ρ exceeds one. In the log-linear consumption dynamics, this adjustment for risk induced by γ is constant. An

important input into this adjustment is the vector

$$H + \beta B'(I - \beta A')^{-1}G. \tag{22}$$

To interpret this object, notice that the impulse response sequence for consumption growth to a shock w_{t+1} is: $H'w_{t+1}$, $G'Bw_{t+1}$, $G'ABw_{t+1}, \ldots$. Then (22) gives the discounted impulse response vector for consumption. It is the variance of this discounted response vector (discounted by β) that enters the constant term of the continuation value as a measure of the risk.

The formulas that follow provide the ingredients for the second-order adjustment in the wealth–consumption ratio and the first-order adjustment in risk adjusted prices.

We use the formula for the continuation value to infer the distorted expectation operator. The contribution of the shock w_{t+1} to $(1-\gamma)v^1_{t+1}$ is given by $(1-\gamma)(H + B'U_v)'w_{t+1}$. Recall that w_{t+1} is a multivariate standard normal. By a familiar complete-the-square argument:

$$\exp\left[(1-\gamma)(H + B'U_v)'w - \frac{1}{2}w'w\right]$$
$$\propto \exp\left(-\frac{1}{2}\big[w - (1-\gamma)(H + B'U_v)\big]'\big[w - (1-\gamma)(H + B'U_v)\big]\right).$$

The left-hand side multiplies the standard normal by the distortion implied by (18) up to scale. The right-hand side is the density of the normal up to scale with mean $(1-\gamma)(H + B'U_v)$ and covariance matrix I. This latter probability distribution is the one used for the distorted expectation operator E^* when computing the derivative of the continuation value. Under this alternative distribution, we may write

$$w_{t+1} = (1-\gamma)(H + B'U_v) + w^*_{t+1}$$

where w^*_{t+1} is a standard normal distribution. As a consequence, consumption and the Markov state evolve as:

$$x_{t+1} = Ax_t + (1-\gamma)B(H + B'U_v) + Bw^*_{t+1},$$
$$c_{t+1} - c_t = G'x_t + \mu_c + (1-\gamma)H'(H + B'U_v) + H'w^*_{t+1}.$$

5.1.5. Example economies

To illustrate the calculations we consider two different specifications of consumption dynamics that include predictable components to consumption growth rates. One of these is extracted from Bansal and Yaron (2004) but specialized to omit time variation in volatility. Later we will explore specifications with time varying volatility after developing a continuous time counterpart to these calculations. This specification is designed to capture properties of consumption variation of the period 1929 to 1998 and is specified at a monthly frequency. The second specification is obtained from an estimation in Hansen, Heaton and Li (2005). In this specification quarterly post World War II data is used. This data is described in Appendix D.

The first specification is:

$$c_{t+1} - c_t = 0.0015 + x_t + [\,0.0078 \quad 0\,]w_{t+1},$$
$$x_{t+1} = 0.98x_t + [\,0 \quad 0.00034\,]w_{t+1}.$$

There are two shocks, one directly impacts on consumption and the second one on the conditional mean of consumption. In the Breeden (1979)–Lucas (1978) specification of preferences with power utility, only the first shock will have a local price that is different from zero. In the recursive utility the second shock will also have a nonzero price because of the role of the continuation value.

Figure 4 reports the impulse response functions for consumption in reaction to the two shocks. The first shock by construction has a significant immediate impact that is permanent. The second shock has a relatively small initial impact on consumption but the effect builds to a significant level. With recursive utility this long-run impact can produce a potentially large effect on risk prices especially since the effect can be magnified by choice of the risk aversion parameter γ.

The second specification is inferred by fitting a vector autoregression of $c_{t+1} - c_t$ and $c_{t+1} - e_{t+1}$ the logarithm of the ratio of consumption to corporate earnings. It is

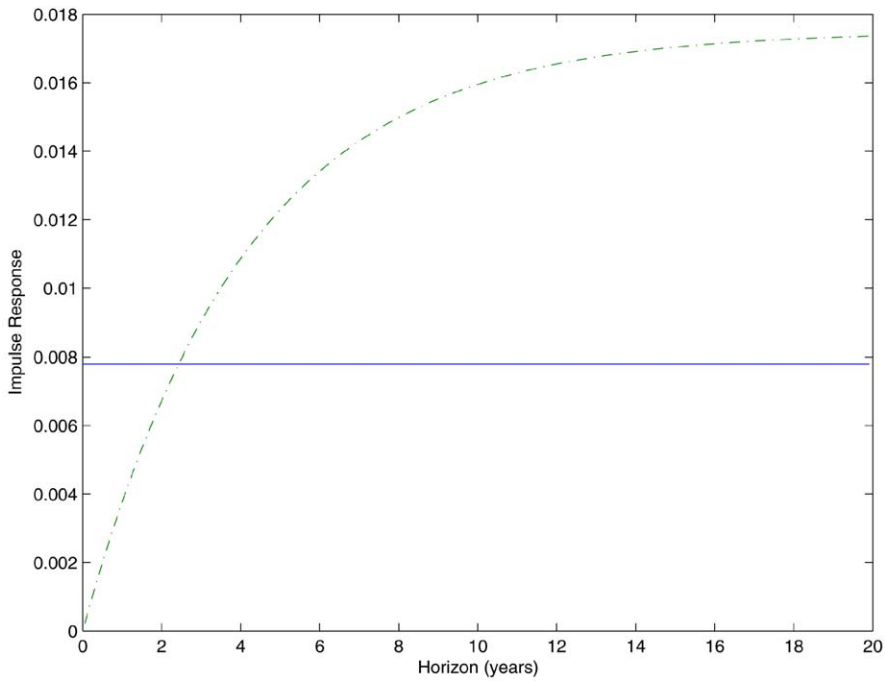

Figure 4. Consumption impulse responses implied by Bansal–Yaron model. — depicts response of consumption to a consumption shock. – · – depicts response of consumption to a predicted consumption shock.

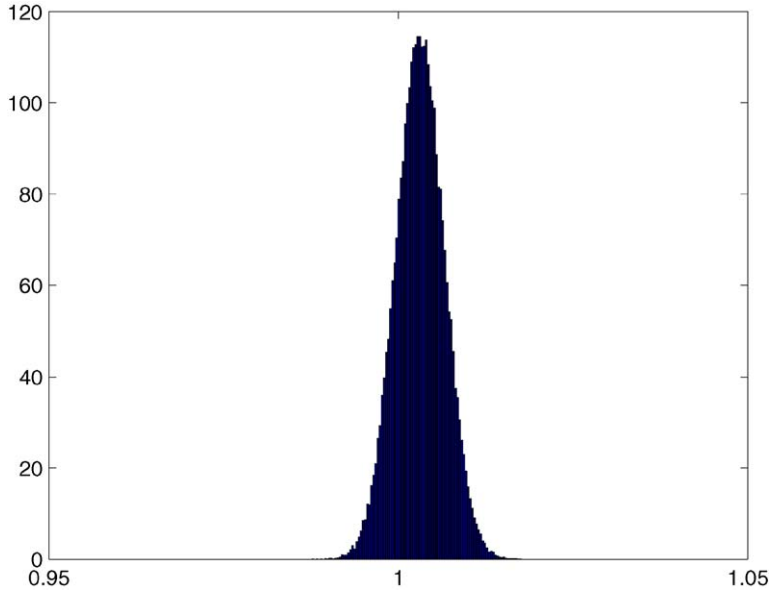

Figure 5. Approximate posterior distribution for cointegration parameter. Construction uses Box–Tiao priors for each equation of the VAR for consumption and corporate earnings. The posterior distribution is for the parameter λ where $c_{t+1} - \lambda e_{t+1}$ is assumed to be stationary. The histogram is scaled to integrate to one.

important in this specification that corporate earnings and consumption are cointegrated with a coefficient of one. Most models of aggregate growth yield this restriction. There is also empirical support for our assumption. For example, consider Figure 5 which reports an approximate Bayesian posterior distribution for the parameter λ where $c_{t+1} - \lambda e_{t+1}$ is assumed to be stationary. This distribution was calculated using the technique described in Appendix B. Notice that the distribution of λ is centered very close to one. There is some variation around this point but it is very minor so that restricting $\lambda = 1$ is empirically grounded.

In this model there are also two shocks. We identify one as being proportional to the one-step ahead forecast error to consumption scaled to have a unit standard deviation. The second shock is uncorrelated with this first shock and has no immediate impact on consumption. Figure 6 reports the estimated response of consumption to the two shocks. Notice that both shocks induce important long-run responses to consumption that are different from the short-run impulse. For example, the long-run response of consumption to its own shock is almost twice the immediate response. As in the Bansal–Yaron model, consumption has an important low-frequency component. With recursive preferences this low-frequency component can have an important impact on risk premia.

We can identify shocks using an alternative normalization that emphasizes long-run effects. In particular we identify one shock from the VAR that has a transient effect with no impact on consumption in the long run. The other shock is uncorrelated with

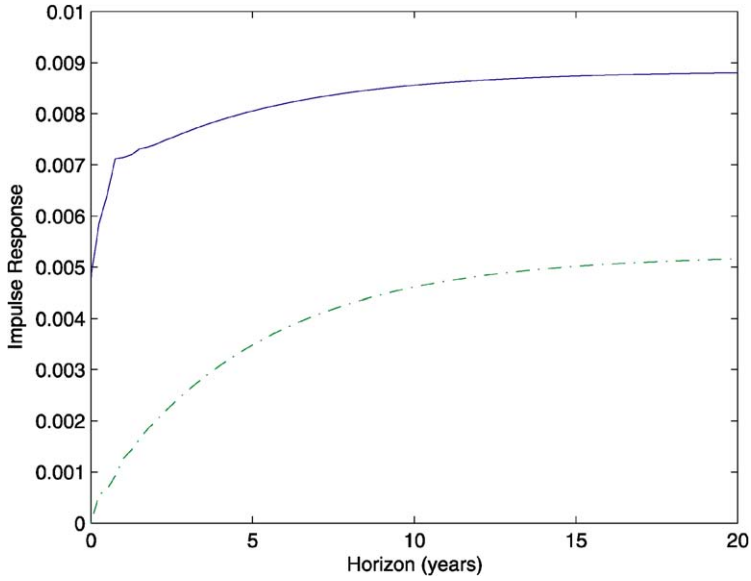

Figure 6. Impulse responses implied by the VAR of Hansen–Heaton–Li model. — depicts response to a consumption shock. – · – depicts response to an earnings shock.

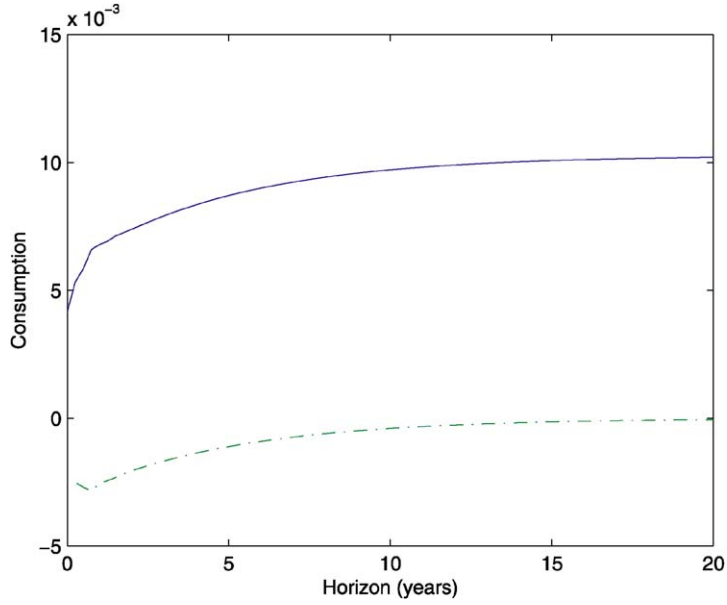

Figure 7. Impulse responses of consumption to permanent and temporary shocks. — depicts impulse response to a permanent shock. – · – depicts impulse response to a temporary shock.

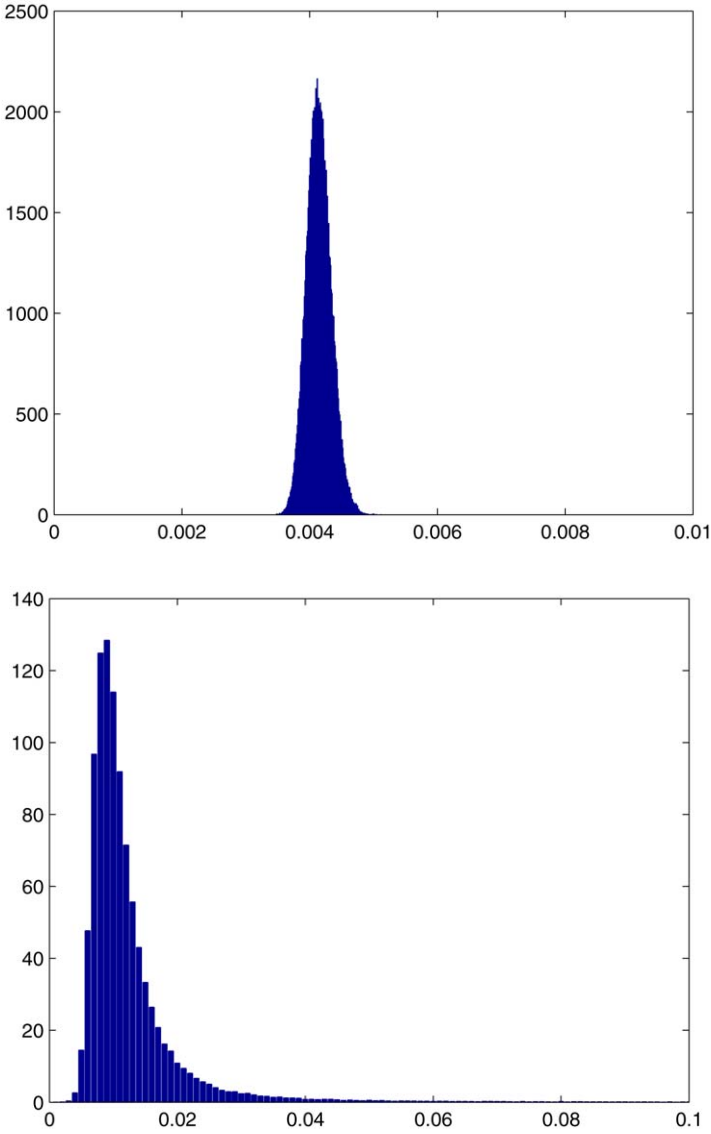

Figure 8. Approximate posterior distributions for responses. The top panel gives the approximate posterior for the immediate response to consumption and the bottom panel the approximate posterior for the long-run response of consumption to the permanent shock. Construction uses Box–Tiao priors for each equation. The histograms are scaled to integrate to one.

this transient shock and has permanent consequences for consumption.[7] The impulse response function of consumption to these two shocks is displayed in Figure 7. Notice that the long-run response to a permanent shock is almost twice the immediate response to this shock.

Although the VAR does identify an important long-run shock to consumption, there is substantial statistical uncertainty surrounding this estimate. To assess this uncertainty we use the technique discussed in Appendix B. Figure 8 reports the approximate posterior distributions for the immediate response of consumption to the temporary shock along with the long-run response of consumption to a permanent shock. Notice that the long-run response is centered at a larger value but that there is uncertainty about this value. The short-run response is measured with much more accuracy.

5.2. Wealth and asset price variation

Pricing models need to imply significant variation in the stochastic discount factor in order to be consistent with some important empirical regularities from financial markets. We also see this when examining aggregate wealth and consumption.

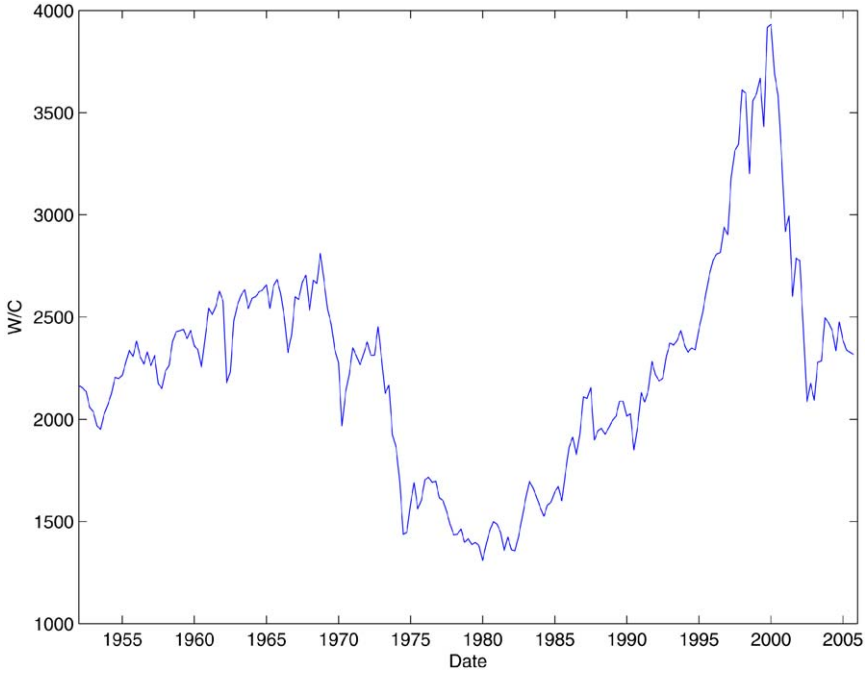

Figure 9. Wealth–consumption ratio from 1952 to 2006.

[7] This approach is an adaptation of the identification scheme advocated by Blanchard and Quah (1989).

When $\rho = 1$ the ratio of consumption to wealth is constant. As we change ρ, this ratio varies. For the alternative models of the dynamics of consumption, we examine whether the pricing model can result in significant variation in the wealth–consumption ratio. This is an important issue because aggregate wealth varies significantly over time due to variation in the market value of wealth. For example in Figure 9 we plot the ratio of wealth to consumption quarterly from 1952 to 2005. Aggregate wealth is measured as the difference between financial wealth and financial liabilities for the household sector of the US economy. This measure of wealth does not include other types of wealth such as human capital.

Notice that there is significant variation in the wealth to consumption ratio. Much of this variation is due to the variability of the market value of traded equity. For example during the late 1990 there was a significant increase in the value of the US stock market which resulted in a substantial increase in the wealth to consumption ratio during this period. With the decline in equity values the wealth to consumption ratio has come back down.

5.2.1. Wealth variation

We now examine the model's implication for wealth when ρ differs from one. We are interested in the effects of alternative parameter values on the predicted level of wealth, the variation in wealth over time and the response of wealth to shocks.

Consider the implications for the wealth–consumption ratio using the dynamics from the VAR with consumption and corporate earnings. Properties of the log wealth–consumption ratio implied by the VAR and the CES model are given in Table 2 for γ and β fixed at 5 and $0.99^{1/4}$ respectively. Several different values of ρ are considered.

Notice that variation in ρ has a significant impact on the forecasted level of the wealth–consumption ratio. Given a value for β this variation could be used to identify ρ based on the observed mean of the ratio. Variation in the mean of the wealth–

Table 2
Properties of the log wealth–consumption ratio

	ρ					
	0.5	0.67	0.9	1.1	1.33	1.5
Mean	9.16	7.78	6.39	5.70	5.50	5.74
STD	0.0092	0.0060	0.0017	0.0017	0.0054	0.0079
STD w/o 2nd order term	0.0086	0.0057	0.0017	0.0017	0.0057	0.0086
Corr. with consumption	0.22	0.22	0.23	−0.23	−0.23	−0.24

Notes. The parameters γ and β are fixed at 5 and $0.99^{1/4}$, respectively. Statistics are calculated via simulation based on a times-series simulation with 60,000 draws of the random vector w_t. The first 10,000 draws were discarded.

consumption ratio induced by ρ can be unwound by choice of β, however. Of interest then is the effect of ρ on the dynamics of the wealth–consumption ratio.

The row "STD" reports the standard deviation of the wealth–consumption ratio which is increasing in the difference between ρ and 1. The row below that ignores term with $(\rho - 1)^2$ in the expansion (19). Notice that this "second-order" term provides little extra variation in the wealth–consumption ratio. Although variation in ρ away from unity does produce variation in the wealth–consumption ratio, this variation is nowhere near the size observed in the data.

The first-order term in the wealth–consumption ratio (19) indicates that shocks to the continuation value affect the wealth–consumption ratio and the sign of the effect depends on the value of ρ relative to 1. In the consumption dynamics estimated by HHL, positive shocks to consumption also have positive impact on the continuation value relative to consumption. When ρ is less than 1 this model predicts a positive covariance between shocks to consumption and wealth. This is reflected in the last line of Table 2 which reports the correlation between the log wealth–consumption ratio and the log consumption growth. Notice that when ρ is less than 1, this correlation is positive. When ρ is greater than 1, this correlation is negative.

To further examine this effect we report the impulse response of the log wealth–consumption ratio with reaction to the two shocks in the VAR in Figure 10. In constructing these impulse response functions we ignored the second-order terms in (19).

Consistent with the correlations between consumption growth and the wealth–consumption ratio reported in Table 2 we see that when ρ is less than 1 a positive shock to consumption has a positive effect on the wealth–consumption ratio. These shocks have positive risk prices in the model and hence a claim on aggregate wealth has a potentially significant risk premium.

The specification considered by Bansal and Yaron (2004) predicts a similar pattern of responses to shocks. Figure 11 reports the response of wealth–consumption ratio to a one standard deviation shock to predicted consumption. Since the first shock has no impact on the state variable the response of wealth–consumption ratio to it is zero in this model. Notice that as in the dynamics estimated by HHL the direction of the response of wealth to a predicted consumption shock depends critically upon the size of ρ relative to unity. When ρ is less than one, the wealth–consumption ratio increases with the shock to predicted consumption. As a result this endogenous price moves positively with consumption and the return on the wealth portfolio is riskier than under the assumption that $\rho = 1.5$.

Since wealth is linked to the continuation value, observed wealth can also be used to identify long-run shocks to consumption. We estimate a bivariate VAR for logarithm consumption growth and the logarithm of the observed wealth–consumption ratio reported in Figure 9. Figure 12 reports the estimated impulse response functions for consumption and wealth implied by this alternative bivariate VAR. As with corporate earnings, the wealth–consumption ratio identifies a potentially important long-run shock to consumption. Notice, however, that the shock to wealth has a very substantial tem-

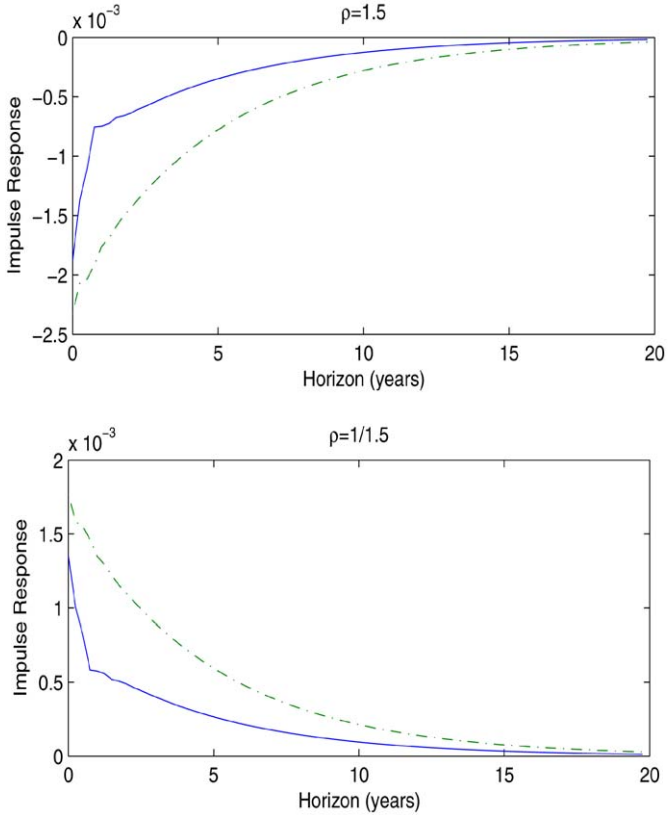

Figure 10. Implied impulse responses of wealth–consumption ratio, Hansen–Heaton–Li model. — depicts impulse response to a consumption shock. – · – depicts impulse response to an earnings shock. The parameters γ and β are set at 5 and $0.99^{1/4}$, respectively.

porary effect on wealth. There is substantial transitory variation in wealth that does not affect consumption as noted by Lettau and Ludvigson (2004).

The relationship between wealth and consumption predicted by the first-order terms of (19) and ρ imposes a joint restriction on the impulse response functions of wealth and consumption. Because of the substantial response of wealth to its own shock, this restriction cannot be satisfied for reasonable values of ρ. As we will see below the necessary variation in ρ results in implausible levels of returns and the wealth–consumption ratio. Ignoring this shock we can examine the restriction of (19) based on the consumption shock alone.

To do this we construct the spectral density of $w_t - c_t - (1-\rho)(v_t^1 - c_t)$ implied by the VAR but setting the variance of the wealth shock to zero. The model implies that at the true value of ρ this density function should be flat. The predicted density is

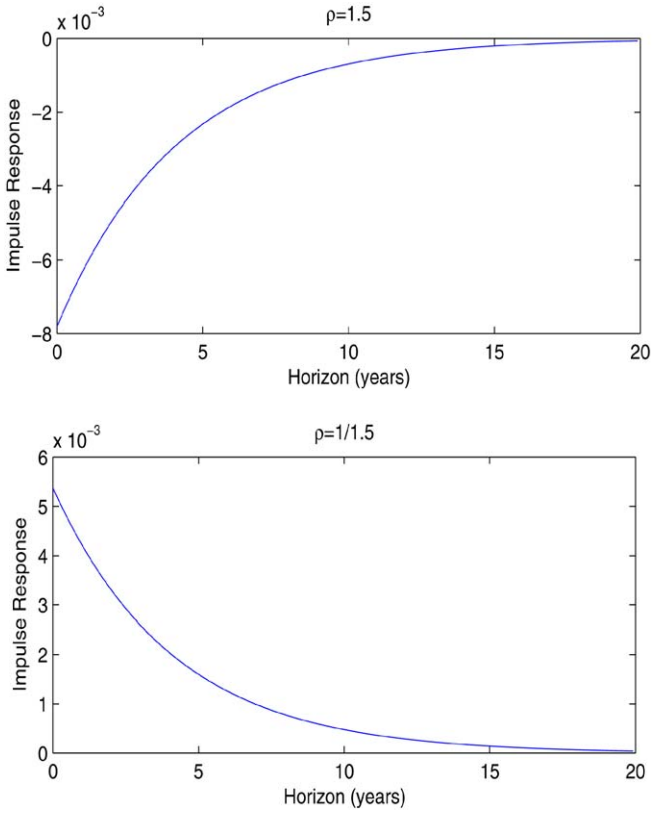

Figure 11. Impulse responses of wealth–consumption ratio to predicted consumption shock, Bansal–Yaron model. The parameters γ and β are set at 5 and 0.998, respectively.

displayed in Figure 13 for $\rho = 0.5$ and $\rho = 1.5$. Smaller values of ρ come closer to satisfying the restriction than the large values of ρ as we will see in Section 7.

5.2.2. Measurement of wealth

Inferences drawn from the recursive utility model based on direct measures of aggregate wealth are sensitive to the wealth proxy used. With a fully specified model of the dynamics of consumption, we circumvent this issue since we can construct implied continuation values and the stochastic discount factors needed to price any series of cash flows. We are therefore able to examine the model's implications for any part of aggregate wealth once we specify the dynamics of the cash flows accruing to the wealth component.

A particularly important part of aggregate wealth is human capital which by its nature is not included in direct measures of wealth. Unobserved human capital may move in

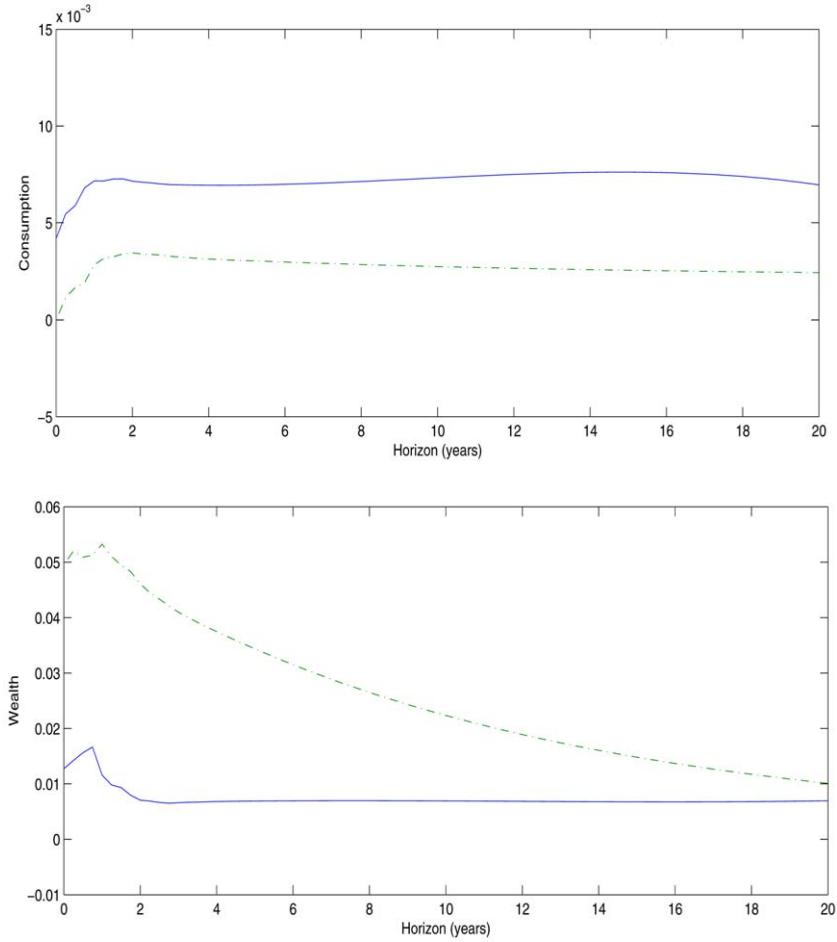

Figure 12. Impulse responses of consumption and wealth. Results from bivariate VAR with consumption growth and the wealth–consumption ratio. — depicts the response to a consumption shock. – · – depicts the response to a wealth shock.

a way that offsets variation in measured wealth so that the true wealth to consumption ratio is relatively constant as predicted by the recursive utility model with ρ close to one. Lustig and Van Nieuwerburgh (2006) use this idea to infer the dynamics of unobserved human capital. As an alternative we specify a dynamic model of the cash flows produced by human capital.

In our analysis we assume that these cash flows are given by labor income. We measure labor income as "Wages and salary disbursements" as reported by the National Income and Product Accounts. As with corporate earnings, we impose the restriction

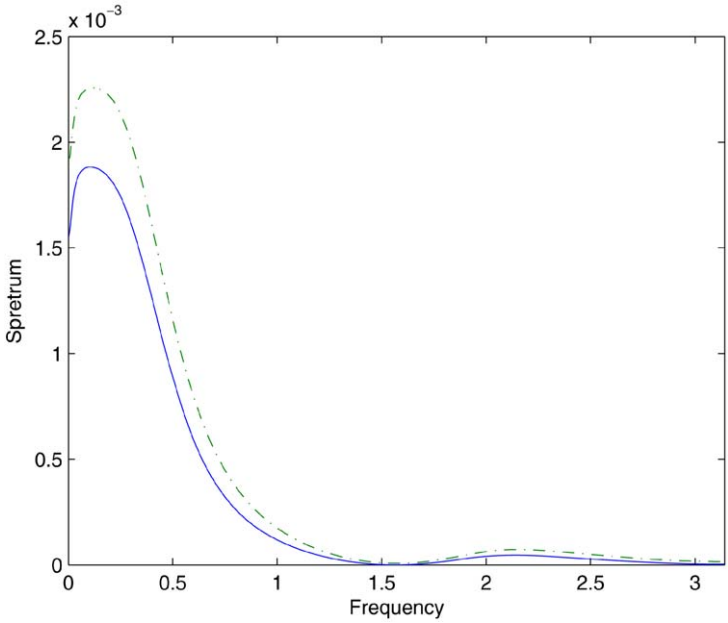

Figure 13. Spectral density of $w_t - c_t - (1-\rho)(v_t^1 - c_t)$. Results are from a bivariate VAR with consumption growth and the wealth–consumption ratio. The variance of wealth shocks is set to zero. — depicts the density for $\rho = 0.5$. – · – depicts the density when $\rho = 1.5$.

Table 3
Summary statistics for corporate and human capital

Capital measure	Standard deviation	Correlation with corporate capital
Human capital	0.056	0.56
Corporate capital	0.033	1
Total	0.034	0.70

Note. Statistics are reported for the natural logarithm of each measure of capital relative to consumption.

that labor income and aggregate consumption are cointegrated with a unit coefficient. We further assume that $\beta = 0.99^{1/4}$, $\gamma = 5$ and $\rho = 1$.

The model's implication for the standard deviation of the (log) ratio of human capital to consumption is reported in Table 3. For comparison the corresponding standard deviation for the ratio of "corporate capital" to consumption is also calculated by valuing the stream of corporate earnings. This measure of wealth does not correspond to any direct measure of the value of capital held by the corporate sector since corporate earnings do not account for investment. Further earnings are reported after payments to

bond holders. Finally in the table "Total" refers to the ratio of the sum of human plus corporate capital to consumption.

Although there are issues of interpretation with these measures of capital, notice that the implied standard deviations are different from zero and that the ratio of human capital to consumption has the greatest variance. In contrast to the analysis of Lustig and Van Nieuwerburgh (2006), human and corporate capital are predicted to be positively correlated. Further, although the model does predict variation in these measures of wealth relative to consumption, the variation is no where near the level depicted in Figure 9. For example, the standard deviation of the logarithm of measured wealth to consumption is 0.24.

This tension is a standard feature of this type of model. Some additional source of variation due to discount factors appears to be necessary to better fit the observed volatility of aggregate wealth and security prices. In the next subsection we add time varying volatility to consumption which provides one potential source of the required variation.

5.3. Continuous time

So far we have seen how predictability in consumption is related to movements in the wealth consumption ratio. The intertemporal substitution parameter is an important ingredient in this relation. In order to permit the risk aversion parameter γ to play a more central role in this time series variation, we consider an extension in which consumption displays stochastic volatility. This volatility gives a source of time-variation in risk premia. To capture this we introduce square root process as a model of volatility and shift our analysis to continuous time. The continuous time formulation we now explore simplifies the analysis of volatility.

Suppose that:

$$\begin{aligned}
\mathrm{d}x_t &= Ax_t\,\mathrm{d}t + \sqrt{z_t}B\,\mathrm{d}W_t,\\
\mathrm{d}z_t &= \overline{A}(z_t - \mu_z)\,\mathrm{d}t + \sqrt{z_t}\overline{B}\,\mathrm{d}\overline{W}_t,\\
\mathrm{d}\log C_t &= G'x_t\,\mathrm{d}t + \mu_c\,\mathrm{d}t + \sqrt{z_t}H'\,\mathrm{d}W_t + \sqrt{z_t}\overline{H}\,\mathrm{d}\overline{W}_t,
\end{aligned} \tag{23}$$

where the matrix A has eigenvalues with real parts that are strictly negative. The process z is scalar and the coefficient $\overline{A}$ is negative. The processes W and $\overline{W}$ are mutually independent standard Brownian motions. The process W can be multivariate and the process $\overline{W}$ is scalar. The volatility process $\{z_t\}$ follows a Feller square root process and $\overline{A}\mu_z + \frac{1}{2}\overline{B}^2 < 0$. In this specification the process $\{z_t\}$ is used to model macroeconomic volatility in an *ad hoc* but convenient manner.

5.3.1. Continuous time Bellman equation

Consider a stochastic evolution for the continuation value of the form:

$$\mathrm{d}\log V_t^{\rho} = \xi_{v,t}^{\rho}\,\mathrm{d}t + \sqrt{z_t}\sigma_{v,t}^{\rho}\,\mathrm{d}W_t + \sqrt{z_t}\bar{\sigma}_{v,t}^{\rho}\,\mathrm{d}\overline{W}_t.$$

For this continuous time diffusion structure, we derive an equation linking the drift $\xi_{v,t}^{\rho}$ with current consumption and continuation values as well as diffusion coefficients.

For this Brownian motion information structure, the continuous time evolution for the continuation value, indexed by ρ, must satisfy:

$$0 = \frac{\delta}{1-\rho}\left[\left(\frac{C_t}{V_t^{\rho}}\right)^{1-\rho} - 1\right] + \xi_{v,t}^{\rho} + z_t\left(\frac{1-\gamma}{2}\right)\left[\sigma_{v,t}^{\rho}\cdot\sigma_{v,t}^{\rho} + \left(\bar{\sigma}_{v,t}^{\rho}\right)^2\right].$$

Heuristically this can be obtained by taking limits of the discrete time recursion (14) as the sample horizon shrinks to zero. The rigorous formulation of recursive preferences in continuous time is given by Duffie and Epstein (1992b).

Thus

$$\xi_{v,t}^{\rho} = \frac{-\delta}{1-\rho}\left[\left(\frac{C_t}{V_t^{\rho}}\right)^{1-\rho} - 1\right] + z_t\left(\frac{\gamma-1}{2}\right)\left[\sigma_{v,t}^{\rho}\cdot\sigma_{v,t}^{\rho} + \left(\bar{\sigma}_{v,t}^{\rho}\right)^2\right].$$

In the special case in which $\rho = 1$, the drift is given by

$$\xi_{v,t}^{1} = \delta\left(v_t^1 - \log C_t\right) + z_t\left(\frac{\gamma-1}{2}\right)\left[\sigma_{v,t}^{1}\cdot\sigma_{v,t}^{1} + \left(\bar{\sigma}_{v,t}^{1}\right)^2\right]. \tag{24}$$

When $\gamma = 1$, the volatility adjustment for the continuation value vanishes and this recursion coincides with the continuation value for preferences with a logarithmic instantaneous utility function. When γ is different from one, there is an adjustment for the volatility of the continuation value. In particular, when γ exceeds one, there is a penalization for big volatility. Typically we are interested in large values of γ to explain the cross section of asset returns.

In what follows we derive the corresponding asset pricing results for a particular endowment economy specified above.[8]

5.3.2. *Value function when $\rho = 1$*

Guess a continuation value of the form

$$v_t^1 = U_v\cdot x_t + \overline{U}_v z_t + c_t + \mu_v$$

where $v_t^1 = \log V_t^1$ as in the discrete-time solution. Thus

$$\begin{aligned} &U_v'Ax + G'x + \overline{U}_v\overline{A}z - \overline{U}_v\overline{A}\mu_z + \mu_c \\ &\quad = \delta U_v'x + \delta\overline{U}_v z + \delta\mu_v + z_t\left(\frac{\gamma-1}{2}\right)\left[\left|U_v'B + H'\right|^2 + (\overline{U}_v\overline{B} + \overline{H})^2\right]. \end{aligned}$$

[8] Asset pricing applications of these preferences are developed by Duffie and Epstein (1992a). They incorporate these preferences into a standard representative agent economy with exogenous asset returns and endogenous consumption in the style of Merton (1973) and Breeden (1979).

Equating coefficients on x gives

$$U_v'A + G' = \delta U_v'$$

or

$$U_v = (\delta I - A')^{-1}G.$$

This formula for U_v is the continuous time analog of our previously derived discrete time formula given in (21).

Equating coefficients in z_t gives the following equation

$$\overline{U}_v\overline{A} = \delta\overline{U}_v + \frac{\gamma - 1}{2}\left[(\overline{U}_v\overline{B} + \overline{H})^2 + \left|U_v'B + H'\right|^2\right]$$

in the unknown coefficient $\overline{U}_v$. This equation can be solved using the quadratic formula, provided that a solution exits. Typically two solutions to this equation exist, and we select the one that is closest to zero. When $\gamma = 1$, $\overline{U}_v = 0$. Large $\overline{B}$ and large values of γ can result in the absence of a solution. On the other hand, shrinking $\overline{B}$ to zero will cause z_t to be very smooth and ensure a solution. The limit can be thought of as giving us the continuous time counterpart to the discrete-time model specified previously in Section 5.1.4.

Consider the special case in which $\overline{H}$ is zero, and suppose that γ exceeds one. Thus there is no immediate impact of the shock $d\overline{W}_t$ on the growth rate of consumption. When solutions exist, they will necessarily be negative because the quadratic function of $\overline{U}_v$ is always positive for all positive values of $\overline{U}_v$. Thus when volatility increases the continuation value declines. The discrete time wealth–consumption expansion (19) in ρ continues to apply in this continuous time environment. Thus when volatility increases the wealth–consumption ratio will increase as well provided that ρ exceeds one, at least for values of ρ local to unity. Conversely, the ratio declines when ρ is less than one.

Finally, the constant term satisfies

$$\mu_c - \overline{U}_z\overline{A}\mu_z = \delta\mu_v$$

which determines μ_v.

For future reference, the local shock exposure of $\mathrm{d}v_t^1$ is

$$\sqrt{z_t}(B'U_v + H)'\,\mathrm{d}W_t + \sqrt{z_t}(\overline{B}\overline{U}_v + \overline{H})\,\mathrm{d}\overline{W}_t.$$

Thus $\sigma_{v,t}^1 = (B'U_v + H)'$ and $\bar{\sigma}_{v,t}^1 = (\overline{U}_v\overline{B} + \overline{H})$.

5.3.3. Derivative with respect to ρ

Next we derive the formula for the derivative of the continuation value with respect to ρ evaluated at one. Our aim is to produce a formula of the form:

$$v_t^\rho \approx v_t^1 + (\rho - 1)Dv_t.$$

The derivative $\{Dv_t\}$ evolves as an Ito process:

$$\mathrm{d}Dv_t = D\xi_{v,t}\,\mathrm{d}t + \sqrt{z_t}D\sigma_t\,\mathrm{d}W_t + \sqrt{z_t}D\bar{\sigma}_t\,\mathrm{d}\overline{W}_t,$$

where $D\xi_{v,t}$ is drift coefficient and $D\sigma_t$ and $D\bar{\sigma}_t$ are the coefficients that govern the shock exposures. We obtain these coefficients by differentiating the corresponding coefficients for the continuation value process with respect to ρ. For instance,

$$D\xi_{v,t} = \left.\frac{\mathrm{d}\xi^{\rho}_{v,t}}{\mathrm{d}\rho}\right|_{\rho=1}.$$

Recall the formula for the drift:

$$\xi^{\rho}_{v,t} = \frac{-\delta}{1-\rho}\left[\left(\frac{C_t}{V_t^{\rho}}\right)^{1-\rho} - 1\right] + z_t\left(\frac{\gamma-1}{2}\right)\left(\sigma^{\rho}_{v,t}\cdot\sigma^{\rho}_{v,t} + \bar{\sigma}^{\rho}_{v,t}\cdot\bar{\sigma}^{\rho}_{v,t}\right).$$

Differentiating with respect to ρ gives

$$D\xi_{v,t} = \delta\frac{(c_t - v_t^1)^2}{2} + \delta Dv_t + z_t(\gamma-1)\left(D\sigma_{v,t}\cdot\sigma^1_{v,t} + D\bar{\sigma}_{v,t}\cdot\bar{\sigma}^1_{v,t}\right). \tag{25}$$

To compute this derivative, as in discrete time it is convenient to use a distorted probability measure. Thus we use

$$\begin{aligned}\mathrm{d}W_t &= \sqrt{z_t}(1-\gamma)\sigma'_{v,t}\,\mathrm{d}t + \mathrm{d}W^*_t,\\ \mathrm{d}\overline{W}_t &= \sqrt{z_t}(1-\gamma)\bar{\sigma}_{v,t}\,\mathrm{d}t + \mathrm{d}\overline{W}^*_t,\end{aligned}$$

where $\{(W^*_t, \overline{W}^*_t)\colon\ t \geqslant 0\}$ is a multivariate Brownian motion. As a consequence, the distorted evolution is

$$\begin{aligned}\mathrm{d}x_t &= Ax_t\,\mathrm{d}t + (1-\gamma)B(B'U_v + H)z_t\,\mathrm{d}t + \sqrt{z_t}B\,\mathrm{d}W^*_t,\\ \mathrm{d}z_t &= \overline{A}(z_t - \mu_z)\,\mathrm{d}t + (1-\gamma)\overline{B}(\overline{B}\overline{U}_v + \overline{H})z_t\,\mathrm{d}t + \sqrt{z_t}\overline{B}\,\mathrm{d}\overline{W}^*_t,\\ \mathrm{d}\log C_t &= G'x_t\,\mathrm{d}t + \mu_c\,\mathrm{d}t + (1-\gamma)H'(B'U_v + H)z_t\,\mathrm{d}t\\ &\quad + (1-\gamma)\overline{H}(\overline{B}\overline{U}_v + \overline{H})z_t\,\mathrm{d}t + \sqrt{z_t}H'\,\mathrm{d}W^*_t + \sqrt{z_t}\overline{H}\,\mathrm{d}\overline{W}^*_t.\end{aligned} \tag{26}$$

Let $\tilde{D\xi}_{v,t}$ denote the resulting distorted drift for the derivative. Then rewrite Equation (25) as

$$\tilde{D\xi}_{v,t} = \delta\frac{(c_t - v_t^1)^2}{2} + \delta Dv_t^1 \tag{27}$$

which can be solved forward as

$$Dv_t^1 = -\frac{\delta}{2}\int_0^{\infty}\exp(-\delta u)E^*\left[\left(c_{t+u} - v^1_{t+u}\right)^2|x_t, z_t\right]\mathrm{d}u.$$

Dv_t^1 is a linear/quadratic function of the composite Markov state (x, z). See Appendix A.2.

5.3.4. Stochastic discount factor

Let s_t^ρ be the logarithm of the continuous time stochastic discount factor for parameter ρ. This stochastic discount factor process encodes discounting for all horizons from the vantage point of time zero. Specifically $\exp(s_t^\rho)$ is discount factor over horizon t and $\exp(s_{t+\tau}^\rho - s_\tau^\rho)$ is the discount factor for horizon t from the vantage point of date τ. Then

$$\begin{aligned}
\mathrm{d}s_t^\rho &= -\delta\,\mathrm{d}t - \rho\,\mathrm{d}c_t + (\rho-\gamma)\Bigg[\mathrm{d}v_t^\rho - \xi_t^\rho\,\mathrm{d}t \\
&\qquad - z_t\left(\frac{1-\gamma}{2}\right)\left(\sigma_{v,t}^\rho\cdot\sigma_{v,t}^\rho + \bar\sigma_{v,t}^\rho\cdot\bar\sigma_{v,t}^\rho\right)\mathrm{d}t\Bigg] \\
&= -\delta\,\mathrm{d}t - \rho\,\mathrm{d}c_t \\
&\qquad + (\rho-\gamma)\Bigg[\sqrt{z_t}\sigma_{v,t}^\rho\,\mathrm{d}W_t + \sqrt{z_t}\bar\sigma_{v,t}^\rho\,\mathrm{d}\overline{W}_t \\
&\qquad - z_t\left(\frac{\rho-\gamma}{2}\right)\left(\sigma_{v,t}^\rho\cdot\sigma_{v,t}^\rho + \bar\sigma_{v,t}^\rho\cdot\bar\sigma_{v,t}^\rho\right)\mathrm{d}t\Bigg].
\end{aligned}$$

Differentiating, we find that the ρ derivative process $\{Ds_t\colon\ t\geqslant 0\}$ evolves as

$$\begin{aligned}
\mathrm{d}Ds_t &= -\mathrm{d}c_t + \Big[\sqrt{z_t}\sigma_{v,t}^1\,\mathrm{d}W_t + \sqrt{z_t}\bar\sigma_{v,t}^1\,\mathrm{d}\overline{W}_t \\
&\qquad - z_t\left(\frac{1-\gamma}{2}\right)\left(\sigma_{v,t}^1\cdot\sigma_1^\rho + \bar\sigma_{v,t}^1\cdot\bar\sigma_{v,t}^1\right)\mathrm{d}t\Big] \\
&\qquad + (1-\gamma)\Big[\sqrt{z_t}D\sigma_{v,t}\,\mathrm{d}W_t + \sqrt{z_t}D\bar\sigma_{v,t}\,\mathrm{d}\overline{W}_t \\
&\qquad - z_t(1-\gamma)\left(D\sigma_{v,t}\cdot\sigma_{v,t}^1 + D\bar\sigma_{v,t}\cdot\bar\sigma_{v,t}^1\right)\mathrm{d}t\Big].
\end{aligned}$$

Thus the ρ approximation is

$$s_t^\rho \approx s_t^1 + (\rho-1)Ds_t$$

with the following contributions to the stochastic evolution of the approximation:

(a) $-\rho\sqrt{z_t}H'$ – Breeden term for exposure to $\mathrm{d}W_t$ risk;
(b) $-\rho\sqrt{z_t}\overline{H}$ – Breeden term for exposure to $\mathrm{d}\overline{W}_t$ risk;
(c) $\sqrt{z_t}(\rho-\gamma)\sigma_{v,t}^1 + \sqrt{z_t}(\rho-1)(1-\gamma)D\sigma_{v,t}$ – recursive utility adjustment for exposure to $\mathrm{d}W_t$ risk;
(d) $\sqrt{z_t}(\rho-\gamma)\bar\sigma_{v,t}^1 + \sqrt{z_t}(\rho-1)(1-\gamma)D\bar\sigma_{v,t}$ – recursive utility adjustment for exposure to $\mathrm{d}\overline{W}_t$ risk.

5.3.5. Risk prices

Of particular interest is the recursive utility adjustment to the Brownian motion risk prices. The ρ approximations are given by the negatives of the values reported in (b) and (c):

(i) $\sqrt{z_t}\rho H' + \sqrt{z_t}(\gamma-\rho)\sigma_{v,t}^1 + \sqrt{z_t}(\rho-1)(\gamma-1)D\sigma_{v,t}$ – risk prices for exposure to $\mathrm{d}W_t$;

(ii) $\sqrt{z_t}\rho\overline{H} + \sqrt{z_t}(\gamma-\rho)\bar{\sigma}_{v,t}^1 + \sqrt{z_t}(\rho-1)(\gamma-1)D\bar{\sigma}_{v,t}$ – risk prices for exposure to $\mathrm{d}\overline{W}_t$.

These prices are quoted in terms of required mean compensation for the corresponding risk exposure. The first vector is the mean compensation for exposure to $\mathrm{d}W_t$ and the second vector is the mean compensation for exposure to $\mathrm{d}\overline{W}_t$.

The risk premia earned by an asset thus consist of a covariance with consumption innovations (multiplied by the intertemporal substitution parameter) and components representing covariance with innovations in the continuation value (weighted by a combination of intertemporal substitution and risk aversion parameters). This characterization is closely related to the two-factor model derived by Duffie and Epstein (1992a), where the second risk term is the covariance with the total market portfolio.

Consider the special case in which $\overline{H}$ is zero. Then under the Breeden model, the volatility shock $\mathrm{d}\overline{W}_t$ has zero price. Under the forward-looking recursive utility model, this shock is priced. For instance, for large γ and ρ close to one, the contribution is approximately $\sqrt{z_t}(\gamma-1)\overline{B}\overline{U}_v$. The recursive utility also amplifies the risk prices for $\mathrm{d}W_t$ risk exposure. For large γ and ρ close to one the prices are approximately $\sqrt{z_t}(\gamma-1)(H' + U_v'B)$, which is the continuous time counterpart to the discounted impulse response function for consumption growth rates. When the importance of volatility becomes arbitrarily small ($\overline{B}$ declines to zero), the volatility state ceases to vary and collapses to μ_z. The predictability in consumption continues to amplify risk prices but the prices cease to vary over time.

Again we consider two specifications. The first is a continuous time version of Bansal and Yaron (2004). In contrast with our discrete time example, but consistent with Bansal and Yaron (2004), we introduce stochastic volatility:

$$\begin{aligned}
\mathrm{d}c_t &= 0.0015\,\mathrm{d}t + x_t\,\mathrm{d}t + \sqrt{z_t}0.0078\,\mathrm{d}W_{1,t},\\
\mathrm{d}x_t &= -0.021x_t\,\mathrm{d}t + \sqrt{z_t}0.00034\,\mathrm{d}W_{2,t},\\
\mathrm{d}z_t &= -0.013(z_t-1)\,\mathrm{d}t + \sqrt{z_t}0.038\,\mathrm{d}\overline{W}_t.
\end{aligned} \tag{28}$$

By construction the volatility process $\{z_t\}$ has a unit mean.

In the Bansal and Yaron (2004) model, risk premia fluctuate. We use a Feller square root process for conditional variances while Bansal and Yaron (2004) used first-order autoregression with normal errors. In our specification, the stationary distribution for conditional variances is in the gamma family and in their specification the distribution is in the normal family. We report the two densities in Figure 14. Our square root specification is by design analytically tractable and it formally restricts variances to be positive.[9] Thus it is more convenient for our purposes to work with a square root process. The two densities are quite similar, and both presume that there are considerable long run fluctuations in volatility.

[9] Negative variances are very unlikely for the parameter values used by Bansal and Yaron (2004). Moreover, in the unlikely event that zero is reached in a continuous time version of their model, one could impose a reflecting barrier.

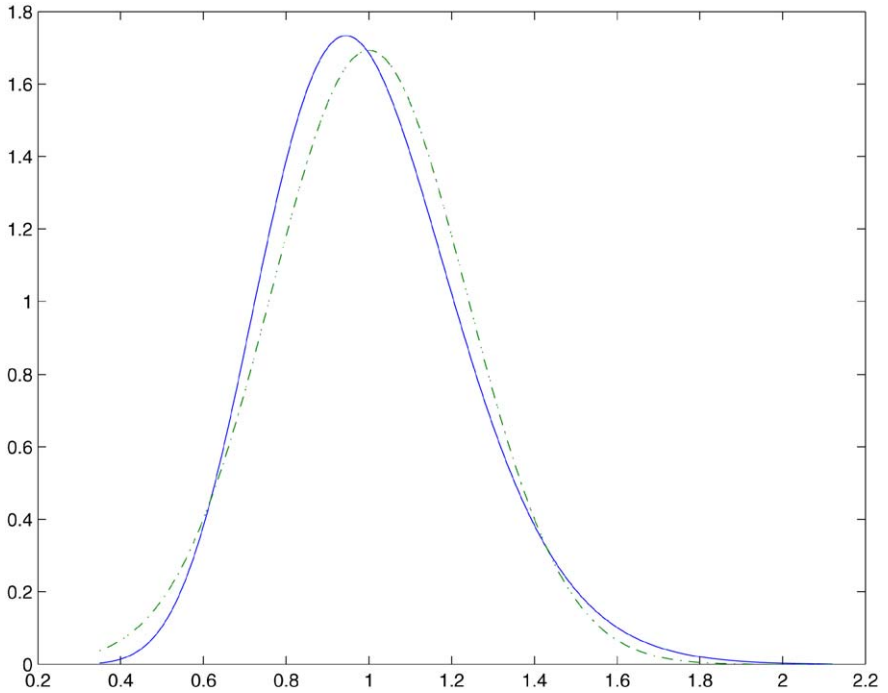

Figure 14. Stationary density of z. — depicts the stationary density of z: gamma(18.0, 0.056). $-\cdot-$ depicts the normal density with the same mean 1 and the same standard deviation 0.236 for comparison.

While we expect γ to have direct impact on risk prices, it is useful to quantify the role of ρ because changing intertemporal substitution parameter will alter risk prices. To quantify this effect, consider the first-order combined expansion in ρ and γ around the values $\rho = 1$ and $\gamma = 1$[10]:

$$\begin{aligned}&\sqrt{z_t}\big[H - (\rho-1)B'U_v + (\gamma-1)(B'U_v + H)\big]\\&\quad= \sqrt{z_t}\left(\begin{bmatrix}2.70\\0\end{bmatrix} - (\rho-1)\begin{bmatrix}0\\5.12\end{bmatrix} + (\gamma-1)\begin{bmatrix}2.70\\5.12\end{bmatrix}\right).\end{aligned}$$

While Bansal and Yaron (2004) use monthly time units, we have rescaled the time units to annual and we have further multiplied prices by one hundred so that the value units are in expected rates of return expressed as percentages.

In contrasting the contributions of ρ and γ, note that while increases in γ amplify both risk prices, increases in ρ reduce the risk price for the shock to the growth rate in consumption. It is the recursive utility adjustment induced by persistence in the growth rate to consumption that makes the risk price of exposure to $\mathrm{d}W_t^2$ different from zero.

[10] This expansion illustrates a point made by Garcia, Renault and Semenov (2006) that when ρ is small, γ underestimates the contribution of risk aversion and conversely when ρ is large.

In this Bansal and Yaron (2004) specification, the risk price of $\mathrm{d}W_t^2$ exposure is double that of $\mathrm{d}W_t^1$. As we will see, the recursive utility contribution is much more challenging to measure reliably.

For pedagogical convenience, we have featured the first-order term in γ, in fact this is not critical. The higher-order term allows us to explore nonlocal changes in the parameter γ. For instance, as we change γ to be five and then ten, the first-order expansions in ρ evaluated at $x_t = 0$ and $z_t = 1$ are:

$$\gamma = 5: \qquad \sqrt{z_t}\left(\begin{bmatrix} 13.5 \\ 20.5 \end{bmatrix} - (\rho - 1)\begin{bmatrix} 0 \\ 5.9 \end{bmatrix}\right),$$

$$\gamma = 10: \qquad \sqrt{z_t}\left(\begin{bmatrix} 27.0 \\ 46.1 \end{bmatrix} - (\rho - 1)\begin{bmatrix} 0 \\ 5.3 \end{bmatrix}\right).$$

The ρ derivatives change as we alter γ, but not dramatically so.

Consider next the price of exposure to volatility risk. For model (28), $\overline{H} = 0$ and the magnitude of $\overline{U}_v$ depends explicitly on the choice of γ. In the local to unity expansion of γ and ρ, level term and the coefficients on both $\rho - 1$ and $\gamma - 1$ are zero suggesting that volatility risk premia are relatively small. When we increase γ we obtain the following first-order expansions in ρ evaluated at $z_t = 1$ and $x_t = 0$:

$$\gamma = 5: \qquad \sqrt{z_t}\big[-2.0 + (\rho - 1)0.7\big],$$

$$\gamma = 10: \qquad \sqrt{z_t}\big[-10.3 + (\rho - 1)1.1\big].$$

The level terms in the risk prices are negative for the volatility shock. While increases in consumption are valued, increases in consumption volatility are not. There is apparently substantial nonlinearity in how these level terms increase in γ. Doubling γ from five to ten leads to a five fold increase in the magnitude of the volatility risk price.

Consider next the continuous time counterpart to our second specification. In this specification there is no stochastic volatility. The first-order expansion in ρ and γ around the values $\rho = 1$ and $\gamma = 1$ is:

$$\begin{aligned}&\big[H - (\rho - 1)B'U_v + (\gamma - 1)(B'U_v + H)\big] \\ &\quad = \left(\begin{bmatrix} 0.96 \\ 0 \end{bmatrix} - (\rho - 1)\begin{bmatrix} 0.79 \\ 1.01 \end{bmatrix} + (\gamma - 1)\begin{bmatrix} 1.75 \\ 1.01 \end{bmatrix}\right).\end{aligned}$$

Again the coefficient on $\rho - 1$ is negative while the coefficient on $\gamma - 1$ is positive so that increasing ρ diminishes the risk prices. The magnitude of the ρ derivative for pricing the shock to corporate earnings is larger than for the shock to consumption, but the reverse is true for the γ derivative. As we change γ to five and then 10, we find that

$$\gamma = 5: \qquad \begin{bmatrix} 7.95 \\ 4.04 \end{bmatrix} - (\rho - 1)\begin{bmatrix} 1.08 \\ 1.63 \end{bmatrix},$$

$$\gamma = 10: \qquad \begin{bmatrix} 16.69 \\ 9.09 \end{bmatrix} - (\rho - 1)\begin{bmatrix} 1.43 \\ 2.36 \end{bmatrix}$$

so the ρ derivatives get larger in magnitude for larger values of γ.

Overall the risk prices are smaller for the second specification than for the first one. Bansal and Yaron (2004) intended to match data going back to 1929 including the pre-war period whereas Hansen, Heaton and Li (2005) used estimates obtained with post-war data. There is much less consumption volatility in this latter sample.

5.3.6. Risk-free rate

Consider next the instantaneous risk-free rate. For an arbitrary ρ, this is given by limit:

$$\begin{aligned} r_{f,t}^{\rho} &= \lim_{\epsilon \downarrow 0} -\log E\big[\exp(s_{t+\epsilon} - s_t)|\mathcal{F}_t\big] \\ &= \delta + \rho G' x_t + \rho \mu_c - \frac{\rho^2 z_t}{2}\big(H'H + \overline{H}^2\big) \\ &\quad + z_t \rho(\rho - \gamma)\big(H' \cdot \sigma_t^{\rho} + H^* \bar{\sigma}_t^{\rho}\big) \\ &\quad - \frac{z_t(\rho - \gamma)(\rho - 1)}{2}\big(\sigma_{v,t}^{\rho} \cdot \sigma_{v,t}^{\rho} + \big(\bar{\sigma}_{v,t}^{\rho}\big)^2\big). \end{aligned} \tag{29}$$

The last two terms on the right-hand side give the contribution for recursive utility and depends in part on the discrepancy between ρ and γ.

In particular, when $\rho = 1$

$$r_{f,t}^{1} = \delta + G' x_t + \mu_c - \frac{z_t}{2}\big(H'H + \overline{H}^2\big) + z_t(1 - \gamma)\big(H' \cdot \sigma_{v,t}^{1} + \overline{H}\bar{\sigma}_{v,t}^{1}\big).$$

The ρ derivative of the risk free rate is

$$\begin{aligned} Dr_{f,t} &= G' x_t + \mu_c + z_t\big[-H' + (2 - \gamma)\sigma_{v,t}^{1} + (1 - \gamma) D\sigma_{v,t}\big] \cdot H' \\ &\quad + z_t\big[-\overline{H} + (2 - \gamma)\bar{\sigma}_{v,t}^{1} + (1 - \gamma) D\bar{\sigma}_{v,t}\big]\overline{H} \\ &\quad - z_t(1 - \gamma)\big(\sigma_{v,t}^{1} \cdot \sigma_{v,t}^{1} + \big(\bar{\sigma}_{v,t}^{1}\big)^2\big). \end{aligned}$$

The approximation is

$$r_{f,t}^{\rho} = r_{f,t}^{1} + (\rho - 1) Dr_{f,t}.$$

While this expression is a bit tedious, it is informative to contrast the local to unity contributions of ρ to those of γ. At $\gamma = 1$, $\bar{\sigma}_{v,t} = 0$ and thus the local approximation is

$$\begin{aligned} &\delta + G' x_t + \mu_c - \frac{z_t}{2}\big(H'H + \overline{H}^2\big) \\ &\quad + (\rho - 1)\big[G' x_t + \mu_c - z_t(H'H + \overline{H}\,\overline{H}) + z_t H' \cdot \sigma_{v,t}^{1} + z_t \overline{H}\bar{\sigma}_{v,t}^{1}\big] \\ &\quad + (\gamma - 1) z_t\big(-H' \cdot \sigma_{v,t}^{1} - \overline{H}\bar{\sigma}_{v,t}^{1}\big). \end{aligned}$$

Importantly, the term multiplying $(\gamma - 1)$ does not include $G' x_t + \mu_c - z_t(H'H + \overline{H}\,\overline{H})$. In particular, the conditional mean in the growth rate of consumption, as reflected in $\mu_c + G' x_t$ contributes only to the ρ derivative. Increases in ρ will unambiguously increase $\rho \mu_c$, making the interest rate larger. This can be offset to some extent by

shrinking δ but only up to the point where $\delta = 0$. This tension is a version of Weil (1989)'s risk free rate puzzle. The term

$$(\rho - \gamma) z_t \left(H' \cdot \sigma_{v,t}^1 + \overline{H} \bar{\sigma}_{v,t}^1 \right)$$

has the interpretation of changing probability measures by adding drift $(\rho - \gamma) z_t \sigma_{v,t}^1$ and $(\rho - \gamma) z_t \bar{\sigma}_{v,t}^1$ to the respective Brownian motions $\mathrm{d}W_t$ and $\mathrm{d}\overline{W}_t$. Changing ρ or γ will, of course, alter this term, but

$$z_t \left(H' \cdot \sigma_{v,t}^1 + \overline{H} \bar{\sigma}_{v,t}^1 \right)$$

is typically smaller than the mean growth rate of consumption.[11] More generally, these risk-free rate approximations give a formal sense in which changes in γ have a much more modest impact on the instantaneous interest rate than changes in ρ and allows us to consider a wide range of values of γ.

5.3.7. *Cash flow returns*

As we have seen, the local evolution of the stochastic discount factor implies a vector of local risk prices. Next we explore cash-flow counterparts, including a limiting notion of an expected rate of return that compensates for exposure to cash flow risk.

Consider a cash flow that can be represented as

$$D_t = G_t f(X_t) D_0$$

where G_t is a stochastic growth process initialized to be one at date zero, D_0 is an initial condition and $f(X_t)$ is a transient component and the process X evolves as a Markov process. For instance, the Markov process X could consist of (x, z) with evolution equation (23). Multiperiod discounting from time i to time j is denoted $S_{i,j}$.

Define the expected rate of return to a cash flow as

$$\frac{1}{t} \log E\left[G_t f(X_t) | \mathcal{F}_0 \right] - \frac{1}{t} \log E\left[S_{0,t} G_t f(X_t) | \mathcal{F}_0 \right].$$

Let the gross return to holding a cash flow over a unit horizon be

$$\log E\left(S_{1,t} G_t f(X_t) | \mathcal{F}_1 \right) - \log E\left(S_{0,t} G_t f(X_t) | \mathcal{F}_0 \right).$$

An equity is a portfolio of claims to such returns. Both of these returns typically have well-defined limits as $t \to \infty$ and these limits will remain invariant over a class of functions f used to define transient components to cash flows. As emphasized by Hansen,

[11] This term is 0.07 (in annualized percent) in the Bansal and Yaron (2004) model, which is small relative to the 1.8 percent growth rate in consumption when evaluated at $z = 1$. In the Hansen, Heaton and Li (2005) model this term is 0.02 percent which is small relative to a per capita consumption growth rate of 2.9 percent. The remaining term from consumption volatility $z_t(H'H + \overline{H}^2)$ at $z = 1$ is also small, 0.07 in the Bansal and Yaron (2004) model and 0.01 in the Hansen, Heaton and Li (2005) model.

Heaton and Li (2005) and Lettau and Wachter (2007), the intertemporal composition of these returns is of interest.

As featured by Hansen, Heaton and Li (2005) and Hansen (2006), we can construct long run counterpart to risk prices by considering the long run excess returns for alternative G specified by martingales that feature the components of cash flow risk. To be concrete, suppose that:

$$\mathrm{d}\log G_t = -\frac{1}{2}(K'K + \overline{K}'\overline{K})z_t + \sqrt{z_t}K'\,\mathrm{d}W_t + \sqrt{z_t}\overline{K}\,\mathrm{d}\overline{W}_t. \tag{30}$$

This specification allows us to focus on the growth rate risk exposure as parameterized by K and $\overline{K}$. For instance, K and $\overline{K}$ can be vectors of zeros except on one entry in which there is a nonzero entry used to feature this specific risk exposure.

Then the logarithm of the limiting cash flow return is

$$\lim_{t\to\infty}\left(\frac{1}{t}\log E\big[G_t f(X_t)|\mathcal{F}_0\big] - \frac{1}{t}\log E\big[S_{0,t}G_t f(X_t)|\mathcal{F}_0\big]\right) = \eta - \nu.$$

The derivative of $\eta - \nu$ with respect to K and $\overline{K}$ gives the long run cash flow counterpart to a local risk price. Using the method of Hansen and Scheinkman (2006), the family of functions f for which these limits remain invariant can be formally characterized. For such functions f, the cash flow contribution $f(X_t)$ can be viewed as *transient* from the vantage point of long run risk prices.

Following Hansen, Heaton and Li (2005), Hansen and Scheinkman (2006) and Hansen (2006), we characterize these limits by solving so-called *principal eigenfunction problems*:

$$\lim_{t\downarrow 0} E\big[G_t\tilde{e}(X_t)|X_0 = X\big] = \eta\tilde{e}(X),$$
$$\lim_{t\downarrow 0} E\big[S_{0,t}G_t\hat{e}(X_t)|X_0 = X\big] = \nu\hat{e}(X).$$

Finally the logarithm of the limiting holding period return is

$$\lim_{t\to\infty}\big[\log E\big(S_{1,t}G_t f(X_t)|\mathcal{F}_1\big) - \log E\big(S_{0,t}G_t f(X_t)|\mathcal{F}_0\big)\big]$$
$$= -\nu + \log\hat{e}(X_1) - \log\hat{e}(X_0) + \log G_1.$$

This latter return has three components: (a) an eigenvalue component, (b) a pure cash flow component and (c) an eigenfunction component. The choice of the transient component $f(X_t)$ typically does not contribute to the value. The valuation implicit in the stochastic discount factor is reflected in both $-\nu$ and $\log\hat{e}(X_1) - \log\hat{e}(X_0)$, but of course not in the cash flow component $\log G_1$. In contrast to the log-linear statistical decompositions of Campbell and Shiller (1988a), the decompositions we just described require an explicit valuation model reflected in a specification of the stochastic discount factor.

Consider first the Bansal and Yaron (2004) model. The risk prices computed as derivative of long-run return with respect to K depends on the values of K. As the baseline

values of K, we use the risk exposure of the consumption and the state variable. At these baseline values, we obtain the following long run risk prices for $\rho = 1$ as we increase γ[12]:

$$\underset{\gamma=1}{\begin{bmatrix}2.70\\5.62\end{bmatrix}} \quad \underset{\gamma=5}{\begin{bmatrix}13.87\\26.85\end{bmatrix}} \quad \underset{\gamma=10}{\begin{bmatrix}30.30\\58.33\end{bmatrix}}$$

where $\beta = 0.998$ is assumed as in Bansal and Yaron (2004). The prices are close to linear in γ but there is nonlinear contribution caused by stochastic volatility, which makes the risk prices more than proportional to γ. Although the second shock has no immediate impact on consumption and hence a zero local risk price, it has long lasting impact on the stochastic discount factor by altering the predicted growth rate in consumption. As expected in Figure 4, it turns out that the long run risk price for this shock is bigger than that for consumption shock.

Consider next the Hansen, Heaton and Li (2005) model. For this model, the risk prices computed as derivatives of long run return with respect to K are insensitive to the baseline choice of K. In other words the component prices are constant as shown by Hansen, Heaton and Li (2005). For this model we report the long run prices for $\rho = 1$ for three different values of γ:

$$\underset{\gamma=1}{\begin{bmatrix}1.77\\1.06\end{bmatrix}} \quad \underset{\gamma=5}{\begin{bmatrix}8.76\\5.10\end{bmatrix}} \quad \underset{\gamma=10}{\begin{bmatrix}17.50\\10.15\end{bmatrix}}.$$

The prices are linear and are approximately proportional to γ and are computed assuming that $\beta = 0.99^{1/4}$ as in Hansen, Heaton and Li (2005). Even when γ and ρ are one, the long run cash flow risk price is positive for the shock to corporate earnings. While the corporate earnings shock is normalized to have no immediate impact on consumption, it will have a long run impact and hence this will show up in the equilibrium risk prices.

We report the derivatives of long-run risk price with respect to ρ for both specifications in Figure 15. Recall that these derivatives were negative for the local prices. As is evident from this figure, for the Bansal and Yaron (2004) model the derivative is positive for low and high values of γ for the shock to growth rate in consumption. The derivative is negative for a range of intermediate values.

These differences between the derivatives for long run and local prices are due to the predictability of consumption. With the predictability of consumption, the permanent

[12] The prices are slightly decreasing in K. At 10 times baseline values of K, they are

$$\underset{\gamma=1}{\begin{bmatrix}2.69\\5.61\end{bmatrix}} \quad \underset{\gamma=5}{\begin{bmatrix}13.54\\26.80\end{bmatrix}} \quad \underset{\gamma=10}{\begin{bmatrix}28.66\\58.04\end{bmatrix}}.$$

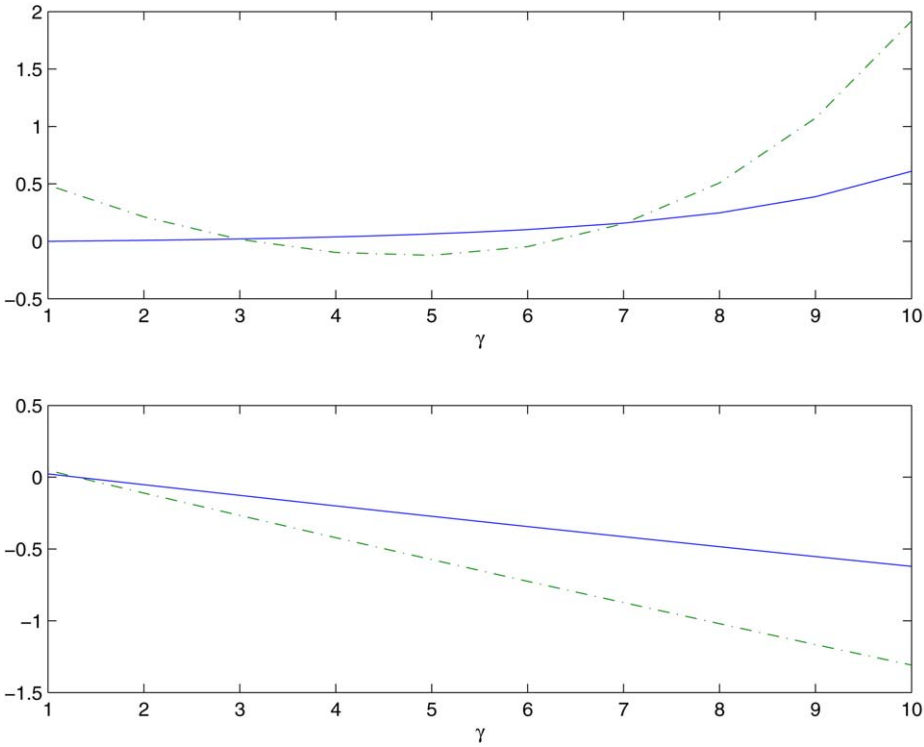

Figure 15. The top panel is Bansal–Yaron model: — depicts ρ derivative of long run risk price of exposure to consumption shock. It is calculated by dividing the difference between ρ derivatives of long-run return at $K = [0\,0]'$ and $K = [0.0078\,0]'$ (risk exposure of c_t) by 0.0078. It is the approximation to the cross derivative of long run return with respect to K and ρ, that is, ρ derivative of long run risk price. The $-\cdot-$ curve depicts ρ derivative of long run risk price of exposure to predicted consumption shock. It is calculated by dividing the difference between ρ derivatives of long-run return at $K = [0\,0]'$ and $K = [0\,0.00034]'$ (risk exposure of x_t) by 0.000034. The bottom panel is Hansen–Heaton–Li model: — depicts ρ derivative of long run risk price of exposure to consumption shock and $-\cdot-$ depicts ρ derivative of the long run risk price of the exposure to corporate earnings. For this model the risk prices, the derivatives with respect to the individual entries of K, are constant.

response of consumption and hence, the permanent response of stochastic discount factor to a shock are more than their contemporary responses. This additional contribution makes the long run risk price and its derivative with respect to ρ larger than their local counterparts. Figure 16 shows this point: long run considerations shift up risk prices and the corresponding ρ derivative.[13]

[13] Because of stochastic volatility, long run considerations tilt the risk price and its derivative along with shifting them.

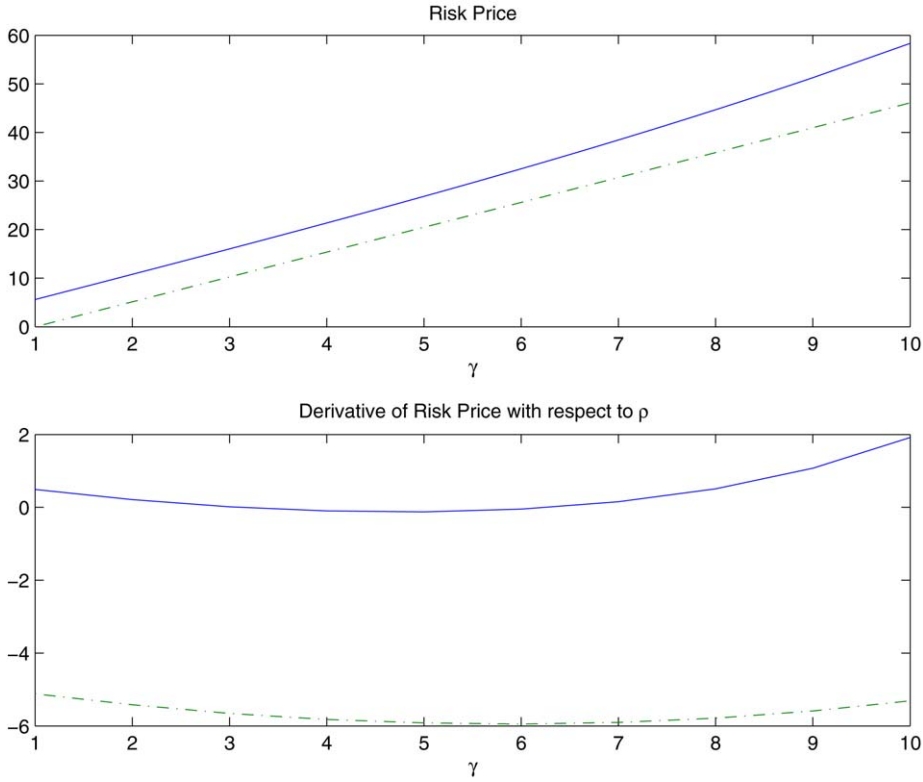

Figure 16. Long run versus local derivatives. Risk price (top panel) and its derivative (bottom panel) with respect to ρ for the shock to growth rate in consumption in Bansal–Yaron model: — depicts long run risk price and ρ derivative; $-\cdot-$ depicts local counterparts. Both levels and derivatives are evaluated at $\rho = 1$.

6. Information about risk aversion from multiple returns

In the previous section we examined how risk aversion and intertemporal substitution affect predicted risk premia. We now examine predictions for risk aversion using information from the returns to the test assets described in Section 4.2. Because of the substantial differences in average returns we will be driven to large levels of risk aversion. For these parameter values, variation in ρ around one has little effect. For this reason and for tractability we assume that $\rho = 1$. For similar reasons Campbell (1996) also considers the case where ρ is close to one and shows that a cross-section of returns can be used to identify γ.

Returns to our test portfolios are known to have differential predictive power for consumption as shown in the work of Parker and Julliard (2005). To the cointegrated model of consumption and corporate earnings of Hansen, Heaton and Li (2005) we add the log

price–dividend ratio and the log dividend growth for each of the five portfolios. To avoid substantial parameter proliferation we estimate each system portfolio by portfolio.

Returning to the discrete time, log-linear setting of Section 5.1, the excess return to an asset is determined by the covariance between shocks to the return and shocks to current and future consumption. As in Section 4 the return to security j has a moving-average representation given by

$$r^j_{t+1} = \rho^j(L)w_{t+1} + \mu^j_r.$$

Hence the on impact effect of the shock vector w_{t+1} on return j is given by the vector $\rho^j(0)$.

Under recursive utility risk premia are determined by the exposure of both consumption and the continuation value to shocks. When the intertemporal elasticity of substitution is assumed to be one, shocks to the log continuation value are given by the discounted impulse responses of log consumption to the shocks. These discounted responses are given by the vector:

$$\Theta(\beta) \equiv H + \beta B'(I - \beta A')^{-1}G.$$

Hence we can write the risk premium for security j as

$$E\big(r^j_{t+1}|\mathcal{F}_t\big) - r^f_{t+1} = -\frac{|\rho^j(0)|^2}{2} + \big[H + (\gamma - 1)\Theta(\beta)\big] \cdot \rho^j(0). \tag{31}$$

Risk aversion can have a large impact on risk premia if consumption is predictable so that $\Theta(\beta)$ is significant and if innovations to discounted future consumption covary with shocks to returns. This covariance is captured by the term $\Theta(\beta) \cdot \rho^j(0)$.

As an initial proxy for this covariance we calculate the covariance between returns at time $t+1$ and $c(t+\tau) - c(t)$ conditional on being at the mean of the state variable and for different values of τ. This calculation ignores discounting through β and truncates the effects at a finite horizon. The results of this calculation are reported in Figure 17 for each of the five book-to-market portfolios. The calculation is done using the point estimates from the VAR for each portfolio.

For small values of τ there is relatively little heterogeneity in the conditional covariance between consumption and portfolio returns. The risk exposure in consumption over the short-term is not a plausible explanation for differences in observed average returns as reported in Table 1. Notice, however, that as τ increases there are pronounced differences in the covariances. For example the covariance between long-run consumption and returns is much higher for portfolio 5 than it is for portfolio 1. Further when $\tau = 40$ the estimated covariances follow the order of the observed average returns. Portfolio 1 has the lowest average return and lowest covariance with consumption. Portfolio 5 has the highest average return and highest covariance.

Figure 18 displays the estimated value of $\Theta(\beta) \cdot \rho^j(0)$ for each security and alternative values of β. As in Figure 17 there are substantial differences in the estimated level of risk exposure across the portfolios as β approaches 1.

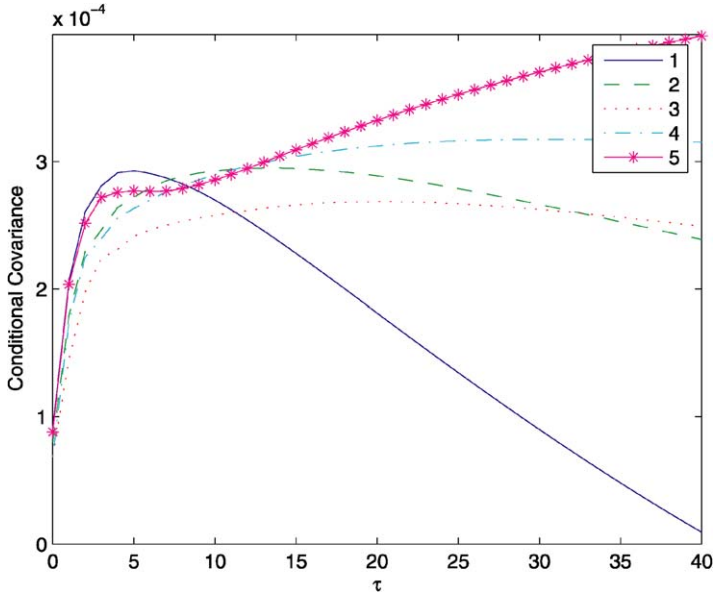

Figure 17. Conditional covariance between returns and future consumption. Conditional covariance between portfolio returns and consumption growth between time t and time $t+\tau$.

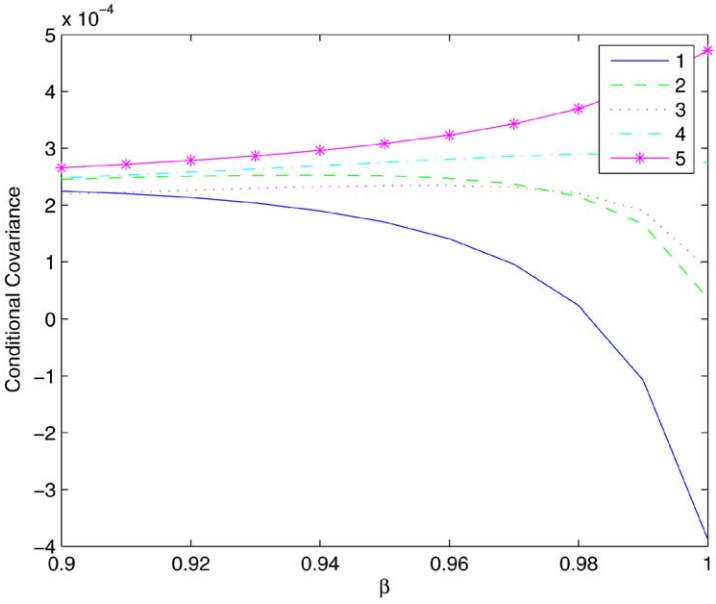

Figure 18. Conditional covariance between returns and $\Theta(\beta)w_{t+1}$.

An implied level of the risk aversion parameter γ can be constructed using the estimates reported in Figure 18. To do this consider the difference between (31) for $j = 5$ and $j = 1$ yields:

$$E\big(r^5_{t+1}|\mathcal{F}_t\big) - E\big(r^1_{t+1}|\mathcal{F}_t\big) = -\frac{|\rho^5(0)|^2}{2} + \frac{|\rho^1(0)|^2}{2} + \big[H + (\gamma - 1)\Theta(\beta)\big] \cdot \big(\rho^5(0) - \rho^1(0)\big).$$

Hence

$$\gamma = \frac{E(r^5_{t+1}|\mathcal{F}_t) - E(r^1_{t+1}|\mathcal{F}_t) + \frac{|\rho^5(0)|^2}{2} - \frac{|\rho^1(0)|^2}{2} - (H - \Theta(\beta)) \cdot (\rho^5(0) - \rho^1(0))}{\Theta(\beta) \cdot (\rho^5(0) - \rho^1(0))}. \tag{32}$$

Using the estimated mean returns reported in Table 1 and the estimates of $\rho^j(0)$ from each VAR system we construct estimates of γ for different values of β. These are given in Table 4. When β is small the estimated value of γ is quite large. Notice, however that as β approaches 1, the two returns have substantially different risk exposures which is reflected in a much smaller estimate of γ.

The estimates reported in Table 4 both ignore sampling uncertainty and are based on estimation that treats each portfolio independently. We repeat the estimation of the VAR except now we consider a six variable system where the dividend growth and price–dividend ratios of portfolio 1 and 5 are included along with $c_t - c_{t-1}$ and $e_t - c_t$. Further we use the Bayesian simulation technique outlined in Appendix B to determine the posterior distribution of the parameters of the VAR systems. For each simulation we infer a value of γ using (32).

In our first set of simulations we ignore the estimation in the mean returns. The quantiles from the posterior distribution of γ are reported in Table 5 where inference about γ

Table 4
Estimates of γ for different values of β, based on (32)

β	γ
0.90	318.1
0.91	252.0
0.92	199.4
0.93	157.0
0.94	122.7
0.95	94.9
0.96	72.2
0.97	53.6
0.98	38.5
0.99	26.1
1.00	16.1

Table 5
Quantiles for γ, mean returns fixed, 5 lags

Quantile:	0.10	0.25	0.50	0.75	0.90
$\beta = 0.98$	−134.66	44.47	76.59	135.94	279.83
$\beta = 0.99$	−58.71	34.53	57.76	99.48	194.87
$\beta = 1$	−14.41	20.72	37.37	63.84	119.84

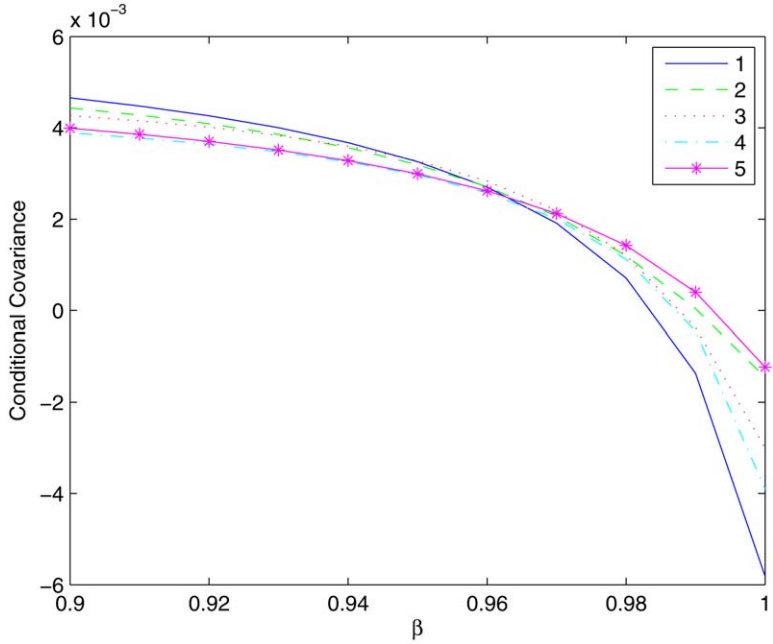

Figure 19. Conditional covariance between returns and $\Theta(\beta)w_{t+1}$. Covariance between shocks to portfolio returns and accumulated shocks to future consumption growth, $\Theta(\beta)w_{t+1}$ for different values of β.

is done conditional on a fixed value of β. Notice that even when β is equal to 1 and sampling error in the means is ignored, there is substantial uncertainty in the estimates of γ.

When $\rho = 1$ the wealth–consumption ratio is constant and innovations in consumption could be measured by innovations to wealth. Since the return on the aggregate wealth portfolio is not observable, a proxy is necessary. A common procedure is to use the return to an aggregate stock index. One justification for this procedure is to assume

that the missing components have returns that are proportional to the stock return as in Campbell (1996).[14]

We repeat the empirical strategy above but assume that the growth rate in consumption is proportional to return on the market portfolio discussed in Appendix D. Figure 19 displays the conditional covariance between the test asset returns and the implied values of $\Theta(\beta)w_{t+1}$ for different values of β. In this case we fit a VAR with 5 lags to the log market return, the log price–dividend ratio for the market along with the log dividend growth and price–dividend ratio for each portfolio.

In this case the implied ordering of risk across the portfolios is consistent with the observed average returns only when β is large enough. When β is small the implied values of γ are negative. For values of β large enough the differences in the covariances between portfolios 1 and 5 imply that portfolio 5 should have a larger return than portfolio 1. Essentially the differential in the return to portfolios 5 and 1, the "return to value" is able to forecast the market return. As in the work of Campbell and Vuolteenaho (2004) the CES model with the market return as a proxy for consumption growth implies that there should be a premium for value over growth: the "value premium".

7. GMM estimation of stochastic discount factor models

For a given financial data set, multiple stochastic discount factors typically exist. Only when the econometrician uses a complete set of security market payoffs will there be a unique discount factor. Either an *ad hoc* identification method is used to construct a discount factor, or an explicit economic model is posed that produces this random variable. Alternative economic models imply alternative measurements of a stochastic discount factor including measurements that depend on unknown parameters. Rational expectations come into play through the use of historical time series data to test relation (2). See Hansen and Singleton (1982) and Hansen, Heaton and Luttmer (1995). Macroeconomics and finance are integrated through the use of dynamic macroeconomic equilibrium models to produce candidate discount factors.

7.1. Identification

As we have seen, pricing restrictions are typically formulated as conditional moment restrictions. For the purposes of this discussion, we rewrite Equation (2):

$$E(S_{t,t+1}a_{t+1}|\mathcal{F}_t) = \pi_t(a_{t+1}) \tag{33}$$

where a_{t+1} is the one period gross payoff to holding an asset. It is a state-contingent claim to the numeraire consumption good at date $t+1$. Suppose an econometrician

[14] Lustig and Van Nieuwerburgh (2006) infer the return to nontraded human capital by using the link between consumption and unobserved wealth implied by several different assumptions about preferences.

observes a vector of asset payoffs: x_{t+1}, a corresponding price vector q_t and a vector of conditioning variables z_t that are measurable with respect to $\mathcal{F}_t$. Moreover, the price vector must be a Borel measurable function of z_t. The vector q_t might well be degenerate and consist of zeros and ones when the payoffs are returns and/or excess returns. An implication of (33) is that

$$E(S_{t,t+1}x_{t+1}|z_t) = q_t. \tag{34}$$

Suppose for the moment that $S_{t,t+1}$ is represented as a nonparametric function of a k-dimensional vector of variables y_{t+1}. That is

$$S_{t,t+1} = f(y_{t+1})$$

for some Borel measurable function f mapping $\mathbb{R}^k \to \mathbb{R}$. Can f be identified? Suppose that we can construct a function h such that h satisfies

$$E\big[h(y_{t+1})x_{t+1}|z_t\big] = 0. \tag{35}$$

Then clearly f cannot be distinguished from $f + \mathsf{r}h$ for any real number r. Thus nonparametric identification depends on whether or not there is a nontrivial solution to (35).

Consider the following problematic examples. If y_{t+1} includes x_{t+1} and z_t, then many solutions exist to (35). For any Borel measurable function g, run a population regression of $g(y_{t+1})$ onto x_{t+1} conditioned on z_t and let $h(y_{t+1})$ be the regression residual:

$$h(y_{t+1}) = g(y_{t+1}) - E\big[g(y_{t+1})x'_{t+1}|z_t\big]\big(E[x_{t+1}x'_{t+1}|z_t]\big)^{-1}x_{t+1}.$$

By construction, this h satisfies (35).

Suppose that we do not impose exclusion restrictions. Instead suppose the vector y_{t+1} includes x_{t+1} and z_t. Stochastic discount factors from explicit economic models are often restricted to be positive. A positive stochastic discount factor can be used to extend the pricing to include derivative claims on the primitive securities without introducing arbitrage.[15] Our construction so far ignores this positivity restriction. As an alternative, we may impose it. Identification remains problematic in this case, there are various ways to construct discount factors.

As shown by Hansen and Jagannathan (1991) and Hansen, Heaton and Luttmer (1995), the solution to the optimization problem

$$\max_{\alpha} -E\big[\big(\max\{-x_{t+1}\cdot\alpha(z_t), 0\}\big)^2|z_t\big] - 2\alpha(z_t)\cdot q_t \tag{36}$$

gives a nonnegative function of x_{t+1} and z_t that solves the pricing equation where α is a function of z_t. From the solution α^* to this concave problem, we may construct a solution to (34) by

$$S_{t,t+1} = \max\{-x_{t+1}\cdot\alpha^*(z_t), 0\}.$$

[15] On the other hand, stochastic discount factors that are negative with positive probability can price incomplete collections of payoffs without inducing arbitrage opportunities.

This is the nonnegative solution that minimizes the second moment. Formally optimization problem (36) is the conjugate to an optimization problem that seeks to find a nonnegative stochastic discount factor that prices the securities correctly whose second moment is as small as possible. Hansen and Jagannathan (1991) were interested in such problems as a device to restrict the set of admissible stochastic discount factors.[16] As demonstrated by Luttmer (1996), convex constraints on portfolios can be incorporated by restricting the choice of α. In contrast to Hansen and Jagannathan (1991), Luttmer (1996) and Hansen, Heaton and Luttmer (1995), we have posed this problem conditionally. We say more about this distinction in the next subsection.

Another extraction choice follows Bansal and Lehmann (1997) and Cochrane (1992) by solving

$$\min_{\alpha} -E\left(\log\left[-\alpha(z_t)\cdot x_{t+1}\right]|z_t\right) - \alpha(z_t)\cdot q_t.$$

Provided this problem has a solution α^*, then

$$S_{t,t+1} = -\frac{1}{\alpha^*(z_t)\cdot x_{t+1}}$$

is a strictly positive solution to (34). This particular solution gives an upper bound on $E[\log S_{t,t+1}|z_t]$. In this case the optimization problem is conjugate to one that seeks to maximize the expected logarithm among the family of stochastic discount factors that price correctly the vector x_{t+1} of asset payoffs.

A variety of other constructions are also possible each of which is an extremal point among the family of stochastic discount factors. Conjugate problems can be constructed for obtaining bounds on convex functions of stochastic discount factors (as in the case of second moments) or concave functions (as in the case of logarithms). As an alternative, Snow (1991) considers bounding other than second moments and Stutzer (1996) constructs discount factors that limit the relative entropy of the implied risk neutral probabilities *vis a vis* the objective probability distribution.

Thus one empirical strategy is to give up on identification and characterize the family of solutions to Equation (34). While this can be a useful way to generate model diagnostics, its outcome for actual pricing can be very limited because the economic inputs are so weak. Alternatively, additional restrictions can be imposed, for example, parametric restrictions or shape restrictions. Motivated by asset pricing models that exhibit habit formation Chen and Ludvigson (2004) specify a stochastic discount factor as a semiparametric function of current and lagged consumption. They use sieve minimum distance estimation in order to identify the shape of this function. In what follows we will focus on parametric restrictions. We consider estimation with parametric restrictions, say $S_{t,t+1} = f(y_{t+1}, \beta)$ for β contained in a parameter space $\mathbb{P}$, a subset of $\mathbb{R}^k$, by fitting the conditional distribution of x_{t+1} and y_{t+1} conditioned on z_t. (As a warning

[16] While this solution need not be strictly positive with probability one, it is nevertheless useful in restricting the family of strictly positive stochastic discount factors.

to the reader, we have recycled the β notation. While β is now a vector of unknown parameters, $\exp(-\delta)$ is reserved for the subjective rate of discount. Also we will use the notation α for a different purpose than in Section 2.)

7.2. *Conditioning information*

Gallant, Hansen and Tauchen (1990) fit conditional distributions parameterized in a flexible way to deduce conditional bounds on stochastic discount factors.[17] Relatedly, Wang (2003) and Roussanov (2005) propose ways of imposing conditional moment restrictions nonparametrically using kernel methods. An alternative is to convert the conditional moment restriction into an unconditional moment restriction by applying the Law of Iterated Expectations:

$$E\big[f(y_{t+1},\beta)x_{t+1} - q_t\big] = 0.$$

A concern might be the loss of information induced by the conditioning down.

As shown by Hansen and Singleton (1982) and Hansen and Richard (1987), this loss can be reduced by expanding the array of assets. For instance consider any vector of conditioning variables $h(z_t)$ with the same dimension as x_{t+1}. Then $x_{t+1}\cdot h(z_t)$ should have a price $h(z_t)\cdot q_t$. Thus it is straightforward to increase the number of asset payoffs and prices by forming synthetic securities with payoffs $h(z_t)\cdot x_{t+1}$ and prices $q_t\cdot h(z_t)$ through scaling by variables in the conditioning information set of investors.

If we perform such a construction for all possible functions of z_t, that is if we verify that

$$E\big[f(y_{t+1},\beta)h(z_t)'x_{t+1} - h(z_t)'q_t\big] = 0$$

for any bounded Borel measurable vector of functions h, then it is necessarily true that

$$E\big[f(y_{t+1},\beta)x_{t+1} - q_t|z_t\big] = 0.$$

This, however, replaces a finite number of conditional moment restrictions with an infinite number of unconditional moment restrictions. It suggests, however, a way to approximate the information available in the conditional moment restrictions through the use of unconditional moment restrictions.

For future reference, let X_{t+1} be the entire vector payoffs including the ones constructed by the econometrician and let Q_t be the corresponding price vector. The corresponding unconditional moment restriction is

$$E\big[f(y_{t+1},\beta)X_{t+1} - Q_t\big] = 0. \tag{37}$$

[17] Cochrane and Hansen (1992) show how to use such estimates to decompose the unconditional volatility of stochastic discount factors into on average conditional variability and unconditional variability in conditional means.

7.3. GMM estimation

In this discussion we work with the ℓ-period extension of (37):

$$E\big[f_\ell(y_{t+\ell}, \beta)X_{t+\ell} - Q_t\big] = 0. \tag{38}$$

The most direct motivation for this is that the data used in the investigation are asset payoffs with a ℓ-period horizon: $f_\ell(y_{t+\ell}, \beta)$. If purchased at date t, their payoff is at date $t + \ell$.[18] Then $f_\ell(y_{t+\ell}, \beta)$ is the ℓ-period stochastic discount factor. For instance, consider Example 3.2. Then

$$f_\ell(y_{t+\ell}, \beta) = \exp(-\delta)\left(\frac{C_{t+\ell}}{C_t}\right)^{-\gamma}$$

where $\beta = (\delta, \gamma)$.

Construct the function

$$\phi_t(\beta) = f_\ell(y_{t+\ell}, \beta)X_{t+\ell} - Q_t,$$

of the unknown parameter vector β. The pricing model implies unconditional moment restriction:

$$E\big[f_\ell(y_{t+\ell}, \beta)X_{t+\ell} - Q_t\big] = 0. \tag{39}$$

Using this as motivation, construct

$$\psi_T(b) = \left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\phi_t(b)\right]' W_T(b)\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\phi_t(b)\right]$$

where the weighting matrix W_t is adjusted to allow for the moving-average structure in error terms:

$$W_T(b) = \left[\mathrm{Cov}_T^0(b) + \sum_{j=1}^{\ell-1}\big(\mathrm{Cov}_T^j(b) + \mathrm{Cov}_T^j(b)'\big)\right]^{-1} \tag{40}$$

where

$$\mathrm{Cov}_T^j(b) \doteq \frac{1}{T}\sum_{t=j+1}^{T}\phi_t(b)\phi_{t-j}(b)'.$$

Then the so-called continuous updating GMM estimator (CU) suggested by Hansen, Heaton and Yaron (1996) is given by

$$b_T = \arg\min_{b\in\mathbb{P}} \psi_T(b),$$

[18] Considerations of aggregation over time leads some researchers to very similar econometric considerations, but only as an approximation. See Hall (1988) and Hansen and Singleton (1996). For a more ambitious attempt to address this issue via numerical simulation see Heaton (1995).

although there are well-known two-step and iterated alternatives. Hansen, Heaton and Yaron (1996) give some comparisons of the approaches.

By construction, the GMM criterion function has close ties to the chi-square distribution. In particular when $b = \beta$, then

$$\psi_T(\beta) \quad \Rightarrow \quad \chi^2(n) \tag{41}$$

where n is the number of moment conditions. As emphasized by Hansen, Heaton and Yaron (1996), this by itself gives a way to conduct inferences about the unknown parameter vector. Construct the set of b's for which $\psi_T(b)$ is less than a threshold value where the threshold value is obtained from the chi-square distribution.[19] Stock and Wright (2000) show formally that such a method accommodates a form of weak identification and leads to robust inference. Alternatively,

$$\psi_T(\beta) - \min_{b\in\mathbb{P}} \psi_T(b) \quad \Rightarrow \quad \chi^2(n-k) \tag{42}$$

where k is number of free parameters. The minimized objective function is itself distributed as a chi-square as shown in Sargan (1958) for the linear simultaneous equations model and by Hansen (1982) for the more general GMM estimation environment. Moreover,

$$\psi_T(\beta) = \left[\psi_T(\beta) - \min_{b\in\mathbb{P}} \psi_T(b)\right] + \left[\min_{b\in\mathbb{P}} \psi_T(b)\right] \tag{43}$$

gives a decomposition of $\psi_T(\beta)$ into two components that are asymptotically independent and each have limiting chi-square distributions.

The limiting chi-square distribution for (42) presumes the local identification condition that matrix

$$E\left[\frac{\partial \phi_t}{\partial b}\bigg|_{b=\beta}\right]$$

has full rank k. When the partial derivative matrix has reduced rank or when one considers a sequence of experiments with limiting singularity, as in the work of Stock and Wright (2000), the limiting chi-square distribution given in (42) is no longer valid. Limit approximation (41) remains valid, however. Kleibergen (2005) suggests an alternative approach to using the latter approximation to conduct inferences. To test a candidate value of β, he constructs a test based directly on the first derivative of the CU-GMM objective function. The limiting distribution has a convenient characterization and leads to an alternative chi-square distribution with degrees of freedom equal to the number of free parameters instead of the number of moment conditions. Interestingly, the test does

[19] Stock and Wright (2000) relate this method to an inversion of the Anderson and Rubin (1949) statistic when specialized to the linear simultaneous equations model.

not require the local identification condition.[20] As discussed in Kleibergen (2005) this approach can be applied to testing restrictions and constructing confidence intervals. Also it can be used to produce an alternative decomposition of (43) that can help to distinguish parameter values for which first-order conditions are approximately satisfied but the underlying moment conditions are not satisfied.

7.4. GMM system estimation

As we have seen, the stochastic discount factor formulation often leads directly to a set of estimation equations, but these are estimation equations for a partially identified model. As an alternative, we add in the remaining components of the model and proceed with a system estimation. One stab at this is given in Hansen and Singleton (1996). The log linear, conditional counterpart to (39) in the case of the power utility model is

$$E\left[-\gamma(\log C_{t+\ell} - \log C_t)\mathbf{1}_m + \log x_{t+\ell}|z_t\right] + \omega - \log q_t = 0 \tag{44}$$

where $\mathbf{1}_m$ is an m-dimensional vector of ones and ω is an m-dimensional vector of constants introduced to compensate for taking logarithms and to capture the subjective rate of discount δ. Here we are abstracting from conditional heteroskedasticity. For simplicity, suppose that q_t is a vector of ones and hence its logarithm is a vector of zeros.

System (44) gives m ℓ-period forecasting equations in $m+1$ variables, the m components of $\log x_{t+\ell}$ and $\log C_{t+\ell} - \log C_t$. Following Hansen and Singleton (1996) we could append an additional forecasting equation and estimate the full system as an $m+1$ dimensional system of ℓ-period forecasting equations. The *reduced form* is a system of forecasting equations for $\log x_{t+1}$ and $\log C_{t+\ell} - \log C_t$ conditioned on z_t:

$$\begin{bmatrix} \log C_{t+\ell} - \log C_t \\ \log x_{t+\ell} \end{bmatrix} = \Pi z_t + \varpi + w_{t+\ell}$$

where

$$E(w_{t+\ell} \otimes z_t) = 0.$$

Then under restriction (44), the matrix Π satisfies

$$[\,-\gamma\mathbf{1}_m \quad I_m\,]\Pi = \mathbf{0}_m \tag{45}$$

where $\mathbf{1}_m$ is an m-dimensional vector of ones, I_m is an m-dimensional identity matrix and $\mathbf{0}_m$ is an m-dimensional vector of zeros.

[20] It requires use of an alternative weighting matrix, one which estimates the spectral density at frequency zero without exploiting the martingale structure implicit in multi-period conditional moment restrictions. For instance, $W_T(b)$ given in formula (40) can be replaced by the weighting matrix estimator of Newey and West (1987). While such an estimator tolerates much more general forms of temporal dependence, its rate of convergence is slower than that of (40). On the other hand, the spectral density estimators are, by construction, positive semidefinite in finite samples.

Notice that (44) also implies the conditional moment restriction:

$$E\big([\,-\gamma\mathbf{1}_m \quad I_m\,]w_{t+\ell}|z_t\big) = 0.$$

Hansen and Singleton (1996) show that even if you impose the stronger condition that

$$E(w_{t+\ell}|z_t) = 0.$$

in estimation, this does not distort the asymptotic inferences for the curvature parameter γ. This means that the reduced-form equation can be estimated as a system GMM estimation, with a weighting matrix constructed so that it does not require a prior or simultaneous estimation of γ. Estimates of γ can be constructed as a restricted reduced-form system. Hansen and Singleton (1996) produce inferences in the analogous ways as for the CU-GMM estimator by constructing confidence sets from a GMM objective function by concentrating all but the parameters of interest.

Notice that if $E[\phi_t(\beta)] = 0$ then it is also true that $E[\Phi(\beta)\phi_t(\beta)] = 0$ where Φ is a function that maps elements of parameter space $\mathbb{P}$ into nonsingular matrices. Thus we may use $\phi_t(b)$ in constructing GMM estimators or $\Phi(b)\phi_t(b)$. For instance in the log-linear power utility model just considered we might divide the moment conditions by $\frac{1}{\gamma}$ and instead estimate $\frac{1}{\gamma}$. Both this restricted reduced form method and the CUE method yield an estimator that is invariant to transformations of this type. The same estimator of the original parameter will be obtained, as is the case in maximum likelihood estimation. This invariance property is not shared by other methods such as two-step methods where a weighting matrix is constructed from an initial consistent estimator. Specifically, it is not satisfied by two-stage least squares when the structural equation to be estimated is over-identified.

7.5. Inference by simulation

The shape of GMM objective, beyond just derivative calculations with respect to parameters, is informative. For low dimensional problems or problems with sufficient linearity, we can depict this function, its level sets, its behavior as we vary one parameter while minimizing out others. For nonlinear problems, an alternative convenient method is to follow Chernozhukov and Hong (2003) by constructing

$$\varphi_T(b) \propto \exp\left[-\frac{1}{2}\psi_T(b)\right]$$

over the set $\mathbb{P}$ provided that this set is a compact subset of $\mathbb{R}^k$ with positive Lebesgue measure.[21] The right-hand side function is scaled so that

$$\int_{\mathbb{P}} \varphi_T(b)\,\mathrm{d}b = 1$$

[21] If $\mathbb{P}$ is not compact, then the objective could be scaled by a weighting function that has finite measure over $\mathbb{P}$.

although there will be no need to compute this scaling factor analytically. The choice of the compact parameter space will be potentially important in applications.

Armed with this construction, we may now use MCMC (Markov chain Monte Carlo) methods to summarize properties of the function φ_T and hence of ψ_T. Appendix D illustrates how to implement MCMC approach. MCMC methods are widely used in making Bayesian inferences, but also can be applied to this problem even though we will use a transformed CU-GMM criterion function instead of a likelihood function.[22] We use the MCMC approach as a way to systematically represent the shape of the GMM objective function via random parameter searches, but we will not attempt to give a Bayesian interpretation of this exercise.

Since $\varphi_T(b)$ may be treated mathematically as a density, we may infer "marginals" for individual components of the parameter vector averaging out the remaining components. This integration step is in contrast to practice of *concentration* producing an objective over a single component of the parameter vector by minimizing the GMM objective over the remaining component for each hypothetical value of the single component. Using the random search embedded in MCMC, approximate level sets can also be inferred.[23] Thus this approach can be used fruitfully in characterizing the behavior of the GMM objective function and offers an attractive alternative to minimization and computing derivatives at minimized values.

7.6. Estimation under misspecification

A feature of the weighting matrix W_T in GMM is that it rewards parameter configurations that imply a large asymptotic covariance matrix. A parameter configuration might look good simply because it is hard to estimate, it is hard to reject statistically. A model specified at the level of a set of moment conditions is in reality only partially specified. Even if we knew the true parameters, we would not know the full time series evolution. If we did, we could form a likelihood function. When combined with a prior distribution over the parameters, we could compute the corresponding posterior distribution; and when combined with a loss function we could produce a parameter estimator that solves a Bayesian decision problem. The GMM estimation is meant to side step the specification of the full model, but at a cost of distancing the inferences from Bayesian methods.

Another way to address this issue is to repose the estimation problem by introducing model misspecification. Instead of aiming to satisfy the moment conditions, suppose we wish to get close to such a specification. This requires a formal statement of what is meant by close, and this choice will alter the population value of the objective. For instance, consider the mean square error objective of minimizing

[22] To make this link, view the function $-\psi_T$ as the log-likelihood and φ_T as the posterior density associated with a uniform prior over the parameter space.

[23] Chernozhukov and Hong (2003) justify estimators of the parameter based on averaging or computing medians instead of minimizing the GMM objective.

$E([f_\ell(y_{t+\ell}, b) - S_{t,t+\ell}]^2)$ by choice of $S_{t,t+\ell}$ subject to

$$E[S_{t,t+\ell} X_{t+\ell} - Q_t] = 0.$$

Since the space of stochastic discount factors $S_{t,t+\ell}$ that satisfies this moment restriction can be infinite dimensional, it is most convenient to work with the conjugate problem, which will need to be solved for each value of b. For fixed b the conjugate problem is a finite-dimensional concave optimization problem. In this case of mean square approximation of the parameterized model to an admissible stochastic discount factor $S_{t,t+\ell}$, we follow Hansen, Heaton and Luttmer (1995) and Hansen and Jagannathan (1997) by using the conjugates problems

$$\min_{b\in\mathbb{P}} \max_{\alpha} E\big[f_\ell(y_{t+\ell}, b)^2 - \big[f_\ell(y_{t+\ell}, b) - \alpha \cdot X_{t+\ell}\big]^2 - 2\alpha' Q_t\big] \tag{46}$$

or

$$\min_{b\in\mathbb{P}} \max_{\alpha} E\big[f_\ell(y_{t+\ell}, b)^2 - \big[\max\big\{f_\ell(y_{t+\ell}, b) - \alpha \cdot X_{t+\ell}, 0\big\}\big]^2 - 2\alpha' Q_t\big] \tag{47}$$

where in both cases the inner problem is concave in α. The second conjugate problem is derived by restricting $S_{t,t+\ell}$ to be nonnegative while the first problem ignores this restriction.

In the case of problem (46), the inner maximization problem is solved by

$$\alpha^*(b) = \big[E(X_{t+\ell} X_{t+\ell}')\big]^{-1} E\big[f_\ell(y_{t+\ell}, b) X_{t+\ell} - Q_t\big]$$

provided that $E(X_{t+\ell} X_{t+\ell}')$ is nonsingular. The concentrated objective function for problem (46) expressed as a function of b is

$$E\big[f_\ell(y_{t+\ell}, b) X_{t+\ell} - Q_t\big]' \big[E(X_{t+\ell} X_{t+\ell}')\big]^{-1} E\big[f_\ell(y_{t+\ell}, b) X_{t+\ell} - Q_t\big],$$

which is the population GMM objective function evaluated using

$$\big[E(X_{t+\ell} X_{t+\ell}')\big]^{-1}$$

as a weighting matrix. Importantly, this matrix does not depend on b. There is no reward for imprecision in estimation.

Alternatively, inner part of problem (47) (optimization over α) does not have such a convenient analytical solution nor does it provide a simple link to GMM estimation, but it is constructed by restricting the admissible stochastic discount factors to be nonnegative. Specifically, the inner problem provides a solution to stochastic discount factor that satisfies the pricing restrictions of the form

$$\max\big\{f_\ell(y_{t+\ell}, b) - \alpha^* \cdot X_{t+\ell}, 0\big\}.$$

The term $\alpha^* \cdot X_{t+\ell}$ is a correction term for misspecification, but is limited so that the resulting stochastic discount factor remains nonnegative.

The sample counterparts to problems (46) and (47) are saddle-point versions of M-estimation problems from the statistics literature instead of GMM estimation problems.

In the sample counterpart problems, the sample average objective function is minimized instead of the population objective function.

Hansen and Jagannathan (1997) show that these two problems can be re-interpreted as ones in which the parameters are chosen to minimize pricing errors over alternative families of payoffs, where pricing errors are measured relative to the square root of the second moment of the payoffs. As a consequence, it is informative to characterize either:

$$\max_{\alpha} E\big(f_\ell(y_{t+\ell}, b)^2 - \big[f_\ell(y_{t+\ell}, b) - \alpha \cdot X_{t+\ell}\big]^2 - 2\alpha' Q_t\big),$$

$$\max_{\alpha} E\big(f_\ell(y_{t+\ell}, b)^2 - \big[\max\big\{f_\ell(y_{t+\ell}, b) - \alpha \cdot X_{t+\ell}, 0\big\}\big]^2 - 2\alpha' Q_t\big)$$

as a function of b to assess model performance for alternative parameter values. Of course other measures of discrepancy between the modeled stochastic discount factor $f_\ell(y_{t+\ell}, b)$ and the stochastic discount factors $S_{t,t+\ell}$ that satisfy pricing restrictions can be employed. Provided the objective is convex in the stochastic discount factor $S_{t,t+\ell}$, we will be led to a conjugate problem that is concave in α, the Lagrange multiplier on the pricing equation.

While we have formulated these as unconditional problems, there are obvious conditional counterparts that use $x_{t+\ell}$ in place $X_{t+\ell}$, q_t in place of Q_t and condition on z_t. Then while α is a function of z_t, the problem can be solved separately for each z_t.

7.7. Intertemporal elasticity estimation

Consider first estimation that features a specific set of assets and other payoffs constructed via scaling. We use the power utility specification and make no attempt to separate risk aversion and intertemporal substitution. Arguably, this is designed to feature estimation of the intertemporal substitution elasticity because by focusing on time series data about a single return, the estimation is not confronting evidence about risk prices. In our first-order expansion of the risk free rate, we saw the impact of both ρ and γ on returns. Arguably the impact of changes in ρ might be more potent than changes in γ, and subsequently we will consider multiple returns and the resulting information about γ. Specifically, we will freely estimate ρ with a single return in this subsection and then estimate γ for fixed alternative values of ρ when we study multiple returns in the Section 7.9.

7.7.1. Treasury bills

Let x_{t+1} be the quarterly return to holding Treasury bills, which has price one by construction. In addition to this return we construct two additional payoffs scaling by consumption ratio between dates t and $t-1$, C_t/C_{t-1} and the date t Treasury bill. Thus there were a total of three moment conditions. Nominal Treasury bill returns were converted to real returns using the consumption deflator. We used per-capita consumption.

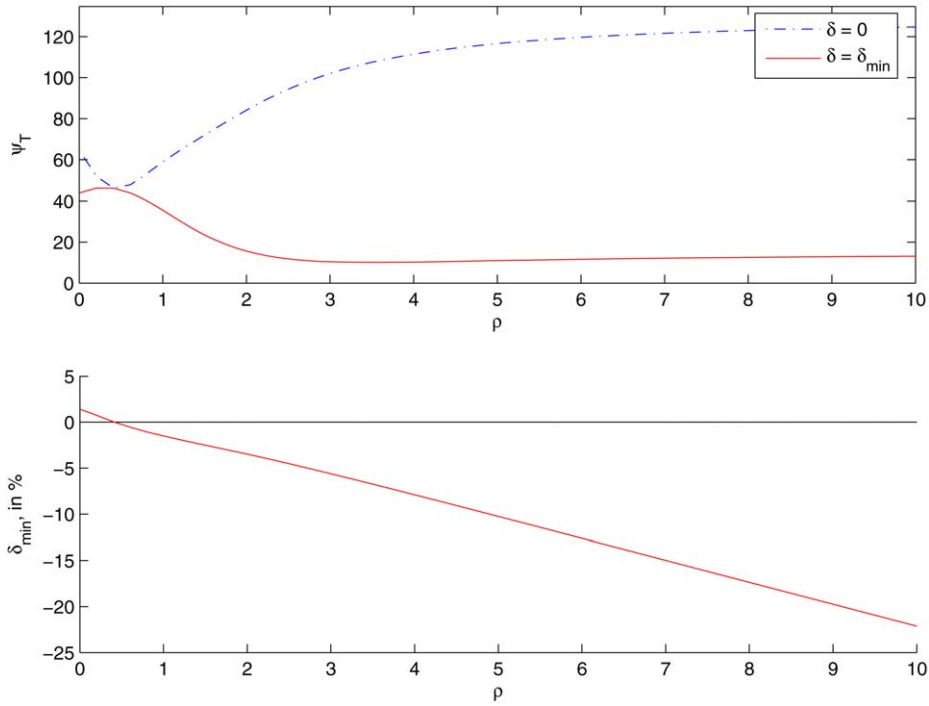

Figure 20. Continuously-updated GMM criterion function for the Treasury bill Euler equation: for $\ell = 1$. The top panel depicts the objective function with and without the constraint that $\delta = 0$. The bottom panel gives the associated values of δ obtained by minimizing the GMM objective for each value of ρ. The parameter δ is expressed as percent per annum.

To facilitate the discussion of inference based on the CU-GMM criterion functions, in Figure 20 we report plots of the concentrated criterion function constructed by minimizing with respect to δ holding ρ fixed over a range of values. We also report the values of the discount rate δ that minimize the criterion concentrated over ρ. The criterion function is minimized at large values of ρ if we do not restrict δ. When we restrict $\delta > 0$, this restriction binds for modest values of ρ and there is notable curvature in the objective function to the right of $\rho = 0.5$. On the other hand, the criterion is very large even at the minimized parameter values. Apparently, it is not possible to satisfy all three moment conditions, even if we allow for sampling uncertainty.

In Figure 21 we construct the payoffs differently. We lag the consumption growth factor and return to Treasury bills one period to remove the effect of overlapping information induced by time aggregation. We also set $\ell = 2$ when constructing the weighting matrix. The shape of the objective (with δ concentrated out), is very similar to that of Figure 20 except that it is shifted down. While reduction in the objective function is to be expected because the conditioning information is less potent, the objective function is

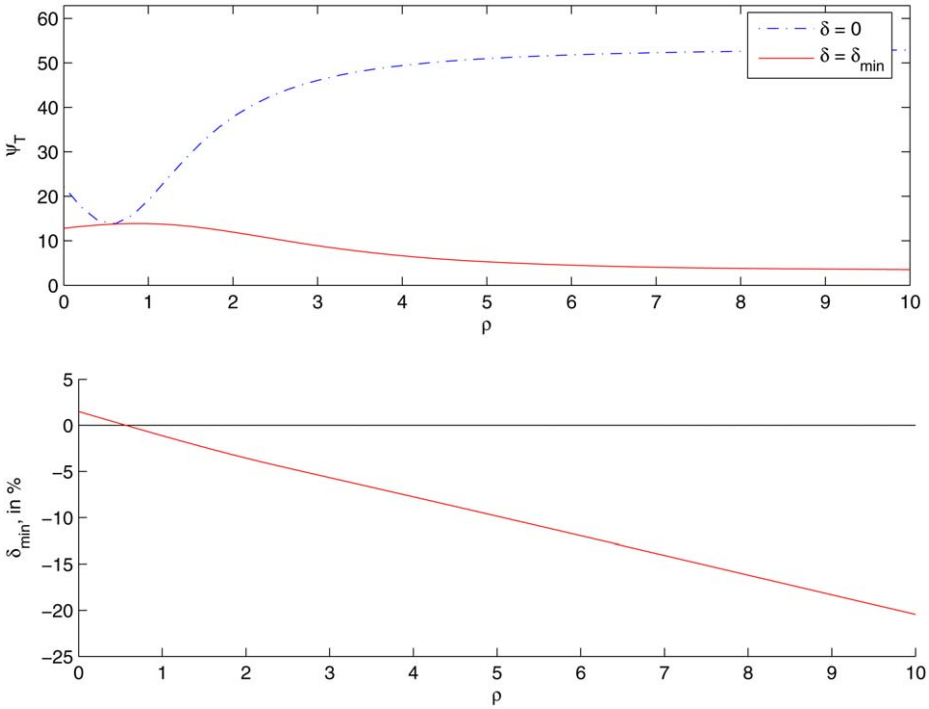

Figure 21. Continuously-updated GMM criterion function for the Treasury bill Euler equation: for $\ell = 2$. The top panel depicts the objective function with and without the constraint that $\delta = 0$. The bottom panel gives the associated values of δ obtained by minimizing the GMM objective for each value of ρ. The parameter δ is expressed as percent per annum.

still quite large. The nonnegativity restriction remains important for inducing curvature to the right of $\rho = 0.5$.

Other researchers have argued that the study of the interest rate Euler equation is fertile territory for weak instrument asymptotics, or more generally for weak formulations of identification.[24] While the evidence for predictability in consumption growth is weak, risk free rates are highly predictable. This is potentially powerful identifying information, suggesting perhaps that the intertemporal elasticity of consumption is very small, ρ is large. Given the observed consumption growth, a large value of ρ requires

[24] Stock and Wright (2000) consider setups in which the expected derivative matrix of the moment conditions drifts to a singular matrix. For the log linear version of the Euler equation, we might ask that the projection of consumption growth onto z_t drifts to zero. If the projection of the Treasury bill onto z_t does not also drift to zero then the coefficient of interest, ρ must drift, changing the nature of the large sample embedding. See Hansen and Singleton (1983) for a related discussion.

a negative subjective rate of discount. Unfortunately, as we have seen this simple argument for large values of ρ ignores restrictions on δ and the overall statistical evidence against the model. Considerations of weak identification are more germane for the study of value-weighted returns.

7.7.2. *Market return*

Next we let x_{t+1} be the value-weighted return. We form two additional payoffs by using consumption growth between date $t-1$ and t along with the date t dividend price ratio. The results are depicted in Figure 22. The objective function is lower than for Treasury bills. Again the imposition of a nonnegativity constraint is inducing curvature in the objective function, in this case to the right of $\rho = 3.5$. For market returns there is

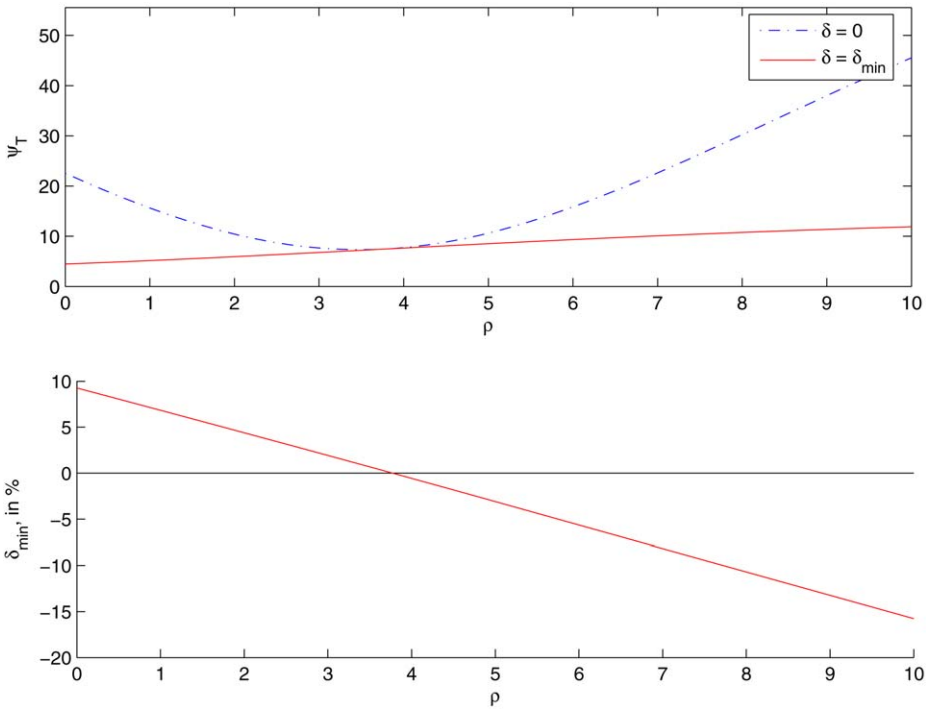

Figure 22. Continuously-updated GMM criterion function for the market return Euler equation: $\ell = 1$. The top panel depicts the objective function with and without the constraint that $\delta = 0$. The bottom panel gives the associated values of δ obtained by minimizing the GMM objective for each value of ρ. The parameter δ is expressed as percent per annum.

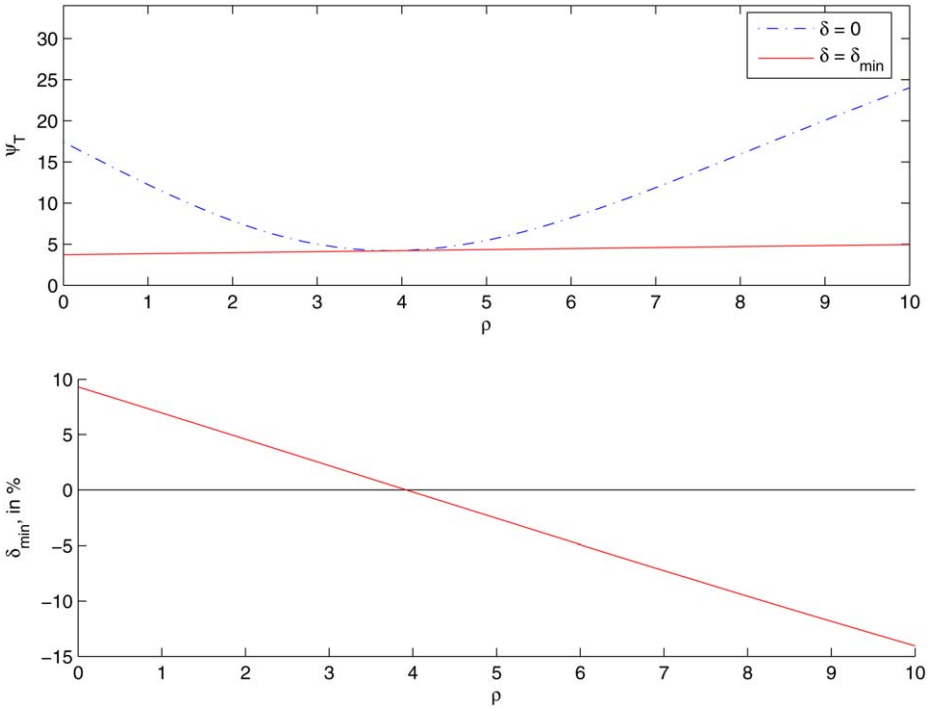

Figure 23. Continuously-updated GMM criterion function for the market return Euler equation: $\ell = 2$. The top panel depicts the objective function with and without the constraint that $\delta = 0$. The bottom panel gives the associated values of δ obtained by minimizing the GMM objective for each value of ρ. The parameter δ is expressed as percent per annum.

considerably less evidence against the model, but also very limited statistical evidence about ρ.[25]

The results when the scaling variable is shifted back one time period are given in Figure 23. Again the shape is similar, and the objective functions is a bit lower.

7.8. CES Preferences and the wealth return

While the CES parameterized version of the recursive utility model gives a leading example of a stochastic discount factor model, as we have seen the stochastic discount factors depend on continuation values. We have already explored constructions of these

[25] The chi-square critical values for two degrees of freedom are 6.0 for probability value of 0.05 and 9.2 for a probability value of 0.01. Since the nonnegativity constraint on δ sometimes binds the chi-square critical values for three degrees of freedom also give a useful reference point. They are 7.8 for probability 0.05 and 11.3 for probability 0.01.

values and their use in empirical investigation. Typically, the computation of continuation values requires a complete specification of the consumption dynamics. In this section we have abstracted from that complication. As emphasized by Epstein and Zin (1989b), an appropriately constructed measure of the wealth return can be used in place of continuation values as we now verify.

Pricing the next period wealth is equivalent to imputing the shadow price to the next period continuation value. Thus we are led to compute

$$\begin{aligned}\frac{E[V_{t+1}MV_{t+1}|\mathcal{F}_t]}{MC_t} &= \left[\frac{\exp(-\delta)}{1-\exp(-\delta)}\right]E\left[(V_{t+1})^{1-\gamma}|\mathcal{F}_t\right]\left[\mathsf{R}(V_{t+1}|\mathcal{F}_t)\right]^{\gamma-\rho}(C_t)^{\rho}\\ &= \left[\frac{\exp(-\delta)}{1-\exp(-\delta)}\right]\left[\mathsf{R}\left(V_{t+1}|\mathcal{F}_t\right)\right]^{1-\rho}(C_t)^{\rho}\end{aligned}$$

where

$$\mathsf{R}(V_{t+1}|\mathcal{F}_t) = \left(E\left[(V_{t+1})^{1-\gamma}|\mathcal{F}_t\right]\right)^{\frac{1}{1-\gamma}}.$$

Thus the return on wealth is given by

$$R^w_{t+1} = \exp(\delta)\left(\frac{C_{t+1}}{C_t}\right)^{\rho}\left[\frac{V_{t+1}}{\mathsf{R}(V_{t+1}|\mathcal{F}_t)}\right]^{1-\rho}.$$

Recall that our previous empirical calculations presumed that $\gamma = \rho$. If we mistakenly impose this restriction, then the Euler equation error is

$$\exp(-\delta)\left(\frac{C_{t+1}}{C_t}\right)^{-\rho}R^w_{t+1} = \left[\frac{V_{t+1}}{\mathsf{R}(V_{t+1}|\mathcal{F}_t)}\right]^{1-\rho}.$$

Suppose that the continuation value is conditionally normally distributed with variance $|\sigma_{v,t}|^2$. While this will typically not be case, it can be justified by taking continuous time limits along the lines we have discussed previously. Then the conditional expectation for this misspecified model is

$$\exp\left[\frac{(1-\rho)(\gamma-\rho)}{2}|\sigma_{v,t}|^2\right].$$

This distortion can be bigger or less than unity depending on whether or not γ is less than or greater than ρ. To the extent that correction is almost constant, it can be absorbed into the subjective rate of discount. Thus GMM estimation with this form of misspecification at the very least alters the restriction imposed on the (potentially distorted) subjective discount rate. Recall that the subjective rate of discount can be an important source of identifying information.

The case of $\gamma = 1$ gives an interesting benchmark. In this special case the log-linear version of the Euler equation holds with:

$$-\delta - \rho\left[\log C_{t+1} - \log C_t\right] + \log R^w_{t+1} = (1-\rho)\left(\log V_{t+1} - E[\log V_{t+1}|\mathcal{F}_t]\right).$$

(See Epstein and Zin (1989b) for an original reference.) In this special case it is not necessary to use the constant term to even approximately correct for volatility in either consumption or the return to wealth. The constant term captures the true subjective rate of discount for investors. Large values of ρ (small values of $\frac{1}{\rho}$) are ruled out by the positive growth rate in per-capita consumption. More generally, studies like those of Hansen and Singleton (1996), and Yogo (2004) report inferences that apparently tolerate large values of ρ, but they ignore restrictions on the constant term. This additional information can be very informative as we have illustrated.[26]

7.9. Multiple assets and Markov chain Monte Carlo

When $\rho \neq 1$, we may invert the relation between continuation values and the return on the wealth portfolio as suggested by Epstein and Zin (1989b):

$$\frac{V_{t+1}}{\mathsf{R}(V_{t+1}|\mathcal{F}_t)} = \left[\exp(-\delta)R_{t+1}^w\right]^{\frac{1}{1-\rho}}\left(\frac{C_{t+1}}{C_t}\right)^{\frac{-\rho}{1-\rho}}.$$

Thus an alternative stochastic discount factor is

$$\begin{aligned} S_{t,t+1} &= \exp(-\delta)\left(\frac{C_{t+1}}{C_t}\right)^{-\rho}\left[\frac{V_{t+1}}{\mathsf{R}(V_{t+1}|\mathcal{F}_t)}\right]^{\rho-\gamma} \\ &= \left[\exp(-\delta)\right]^{\frac{1-\gamma}{1-\rho}}\left(\frac{C_{t+1}}{C_t}\right)^{\frac{\rho(\gamma-1)}{1-\rho}}\left(R_{t+1}^w\right)^{\frac{\rho-\gamma}{1-\rho}}. \end{aligned} \tag{48}$$

The Euler equation for a vector X_{t+1} of asset payoffs with corresponding price vector Q_t is

$$E\left(\left[\exp(-\delta)\right]^{\frac{1-\gamma}{1-\rho}}\left(\frac{C_{t+1}}{C_t}\right)^{\frac{\rho(\gamma-1)}{1-\rho}}\left(R_{t+1}^w\right)^{\frac{\rho-\gamma}{1-\rho}}X_{t+1} - Q_t|z_t\right) = 1$$

where R_{t+1}^w is the return on the total wealth portfolio.

In the empirical analysis that follows, we follow Epstein and Zin (1989b) by using the market return as a proxy for the return on the wealth portfolio. Since the market return omits important components to investor wealth, there are well-known defects in this approach that we will not explore here. Also, we impose some severe restrictions on ρ as a device to illustrate the information available for identifying γ and δ. Freely estimating ρ is problematic because of the poor behavior of the CU-GMM objective in the vicinity of $\rho = 1$. This poor behavior is a consequence of our using an empirical proxy for the return on the wealth portfolio in constructing the stochastic discount factor.

[26] On the other hand, the notion of using single returns to identify ρ independently of γ is typically compromised. The value of γ determines in part what the distortion is in the subjective rate of discount induced by omitting continuation values from the analysis.

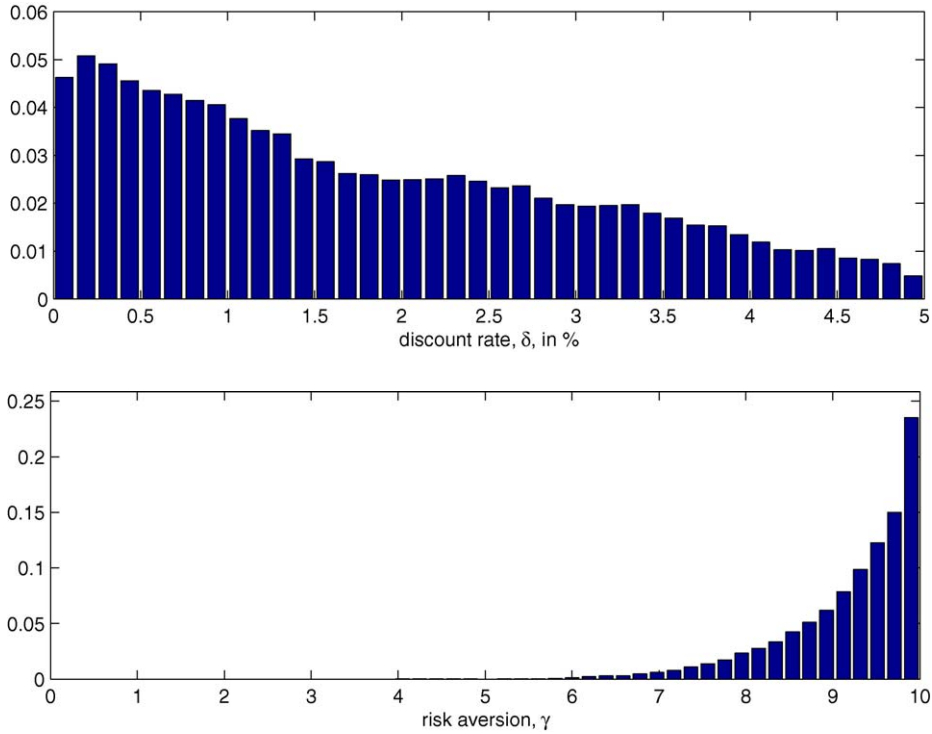

Figure 24. MCMC with the continuously-updated GMM criterion function: $\rho = .5$. The histograms are scaled to integrate to one. The parameter δ is restricted to be in the interval $[0, 5]$ expressed as an annualized percent, and the parameter γ is restricted to be in the interval $[0, 10]$. The smallest CU-GMM objective encountered in the random search was 9.8.

We apply the MCMC simulation method described previously to estimate γ and δ for alternative choices of ρ. This gives us a convenient way to summarize the shape of the CU-GMM criterion function through the use of simulation instead of local approximation. A consequence of our stochastic discount factor construction is that the market portfolio cannot be used as one of the test assets and $\rho = 1$ cannot be entertained. Instead we use the "value minus growth" excess return constructed using the portfolios sorted on book-to-market equity, together with Treasury bill return, in order to identify the preference parameters. The scaling factor for the Treasury bill return are the same ones we used previously, the consumption growth factor between $t - 1$ and t and the time t Treasury bill return. The value-growth excess return is scaled by the consumption growth factor and the date t value-growth excess return. Thus we use six moments conditions in estimation.

In our estimation we use two different values of ρ, $\rho = 0.5$ and $\rho = 1.5$ and estimate γ and δ subject to the constraints that $0 \leqslant \delta \leqslant 5$ and $0 \leqslant \gamma \leqslant 10$ where δ scaled

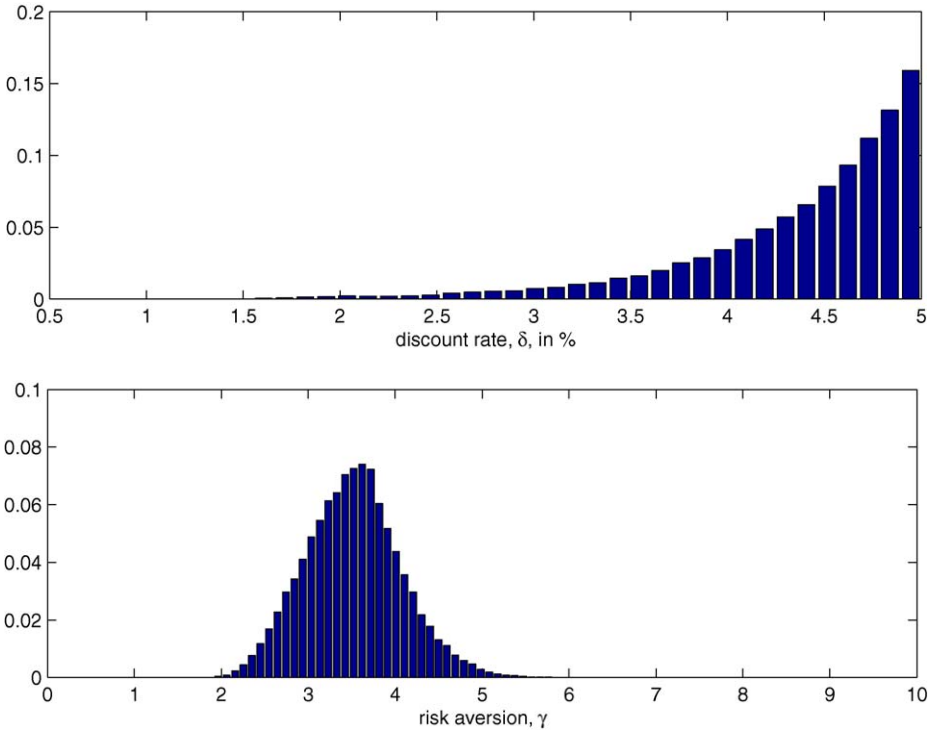

Figure 25. MCMC with continuously-updated GMM criterion function: $\rho = 1.5$. The histograms are scaled to integrate to one. The parameter δ is restricted to be in the interval [0, 5] expressed as an annualized percent, and the parameter γ is restricted to be in the interval [0, 10]. The smallest CU-GMM criterion function value that was encountered in the random search is 21.7.

by 400 so that it is expressed as a percent per annum. The resulting histograms are reported in Figures 24 and 25. When $\rho = 0.5$, the histogram for δ is very much tilted toward zero, and the histogram for γ is very much tilted towards ten. The parameter space bounds play an important role in these calculations, but it is straightforward to impose other bounds. When $\rho = 1.5$, the histogram for γ is centered around 3.5, but the histogram for δ is very much tilted towards the upper bound of five. Increasing the upper bound on δ causes the γ distribution to shift to the right. Thus our chosen upper bound on δ induces a modest estimate of γ. The lowest CU-GMM objective encountered in the random search is 9.8 for $\rho = 0.5$ and 21.7 for $\rho = 1.5$ suggesting that there is considerably less evidence against the specification with a lower value of ρ.[27]

[27] As a point of reference, the critical values for the chi-square distribution with 4 degrees of freedom are 9.5 for a probability value of 0.05 and 13.3 for a probability value of 0.01. Given the important role of the constraints on parameters, the chi-square distribution with five degrees of freedom gives an alternative

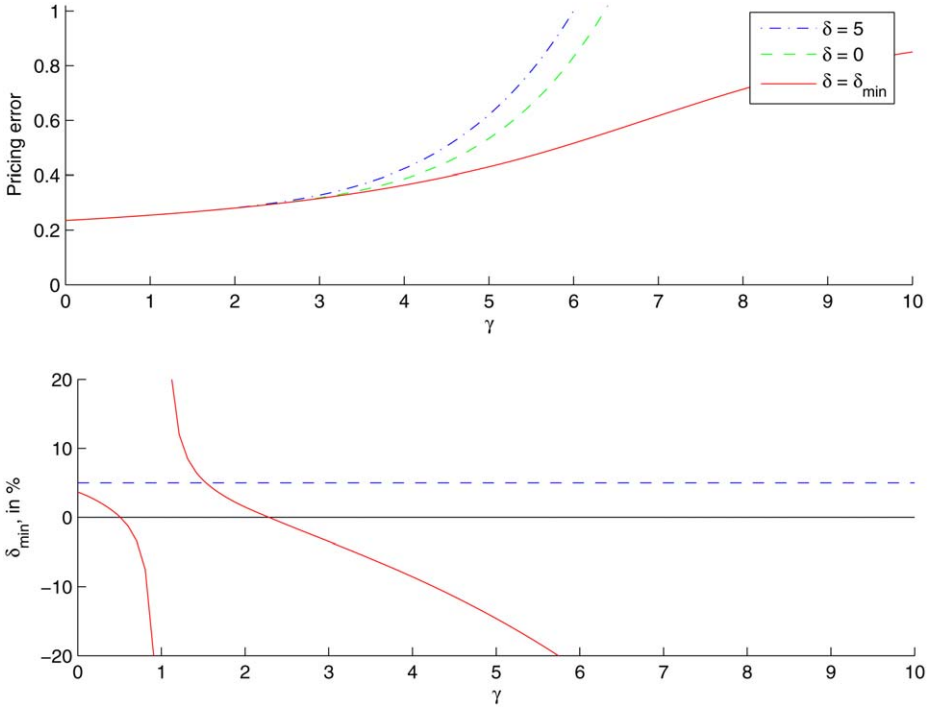

Figure 26. Specification errors: $\rho = 0.5$. The top panel gives the specification error as a function of γ when the value of δ is chosen to minimize the pricing error objective. This pricing error is expressed as the mean-square distance from the misspecified stochastic discount factor to the closest random variable that prices on average the vector of assets. Alternatively, it is the maximal average pricing error per mean-square unit of payoff norm. The bottom panel gives the minimizing choices of δ for each value of γ.

The CU-GMM criterion function has the property that the parameter configurations that induce considerable sampling uncertainty in the moment conditions are tolerated because the weighting matrix is the inverse of the sample covariance matrix. For instance, large values of γ may induce large pricing errors but nevertheless be tolerated. To explore this possibility, we compute the implied specification errors using the weighting matrix described previously. This weighting matrix is invariant to the parameters and instead comes from a best least squares fit of a misspecified model. The outcome of this calculation is depicted in Figure 26 for $\rho = 0.5$ and in Figure 27 for $\rho = 1.5$. When $\rho = 0.5$, the lower bound of zero on δ binds, and the specification errors become large for large values of γ. When $\rho = 1.5$, the upper bound of five binds

interesting benchmark. The critical values are 11.1 for a probability value of 0.05 and 15.1 for a probability value of 0.01.

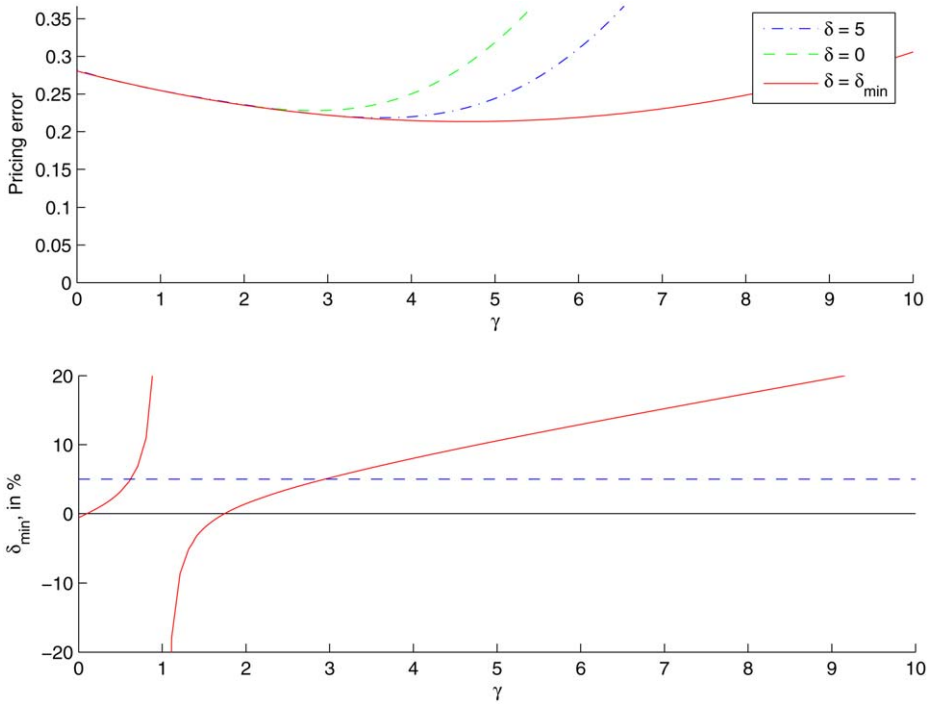

Figure 27. Specification errors: $\rho = 0.5$. The top panel gives the specification error as a function of γ when the value of δ is chosen to minimize the pricing error objective. This pricing error is expressed as the mean-square distance from the misspecified stochastic discount factor to the closest random variable that prices on average the vector of assets. Alternatively, it is the maximal average pricing error per mean-square unit of payoff norm. The bottom panel gives the minimizing choices of δ for each value of γ.

for large values of γ which in turn leads to large specification errors. For both figures the implied value of δ when γ is near one becomes enormous to offset the fact that the subjective discount factor is being raised to a very small number.

8. Conclusions

Our chapter explores the role of intertemporal substitution and risk aversion in asset pricing. We feature the CES recursive utility model, but of course other asset pricing models warrant comparable consideration. Parameters extracted from other sources, including micro or experimental evidence can be inputs into an analysis of the asset pricing implications of models. For example, Malloy, Moskowitz and Vissing-Jorgensen (2005) use evidence from household level data to explore macroeconomic risk. Even with known preference parameters, measurements of macroeconomic risk exposures

are required for quantitative prediction. Since the intertemporal composition of risk can play a central role in asset valuation, this puts an extra premium on the measurement of long-run components to risk. We have not embarked on a comprehensive survey of the empirical literature, but we now explore some of the challenges.

The parameter governing the intertemporal elasticity of substitution is key for linking consumption and wealth. For this link we find it useful to feature the role of continuation values. Since the CES aggregator is homogeneous of degree one, these continuation values encode the shadow values of wealth. In effect the continuation values appropriately scaled give us one side of the intertemporal budget constraint and direct measures of wealth the other side. There is a return counterpart to this link that has been featured in some portions of the asset pricing literature, but the return based formulations typically omit information, in particular information linking current responses of consumption and wealth.

As we have illustrated following the work Lettau and Ludvigson (2001), use of consumption and financial wealth leads to a macroeconomic version of Shiller (1981)'s excess sensitivity puzzle. There is substantial variability in financial wealth that is not reflected in aggregate consumption. This opens up a variety of measurement challenges that have been explored in the asset pricing literature. For example, financial wealth omits any contribution of labor income [see Campbell (1996) and Jagannathan and Wang (1996) for studies of implications for pricing returns], but the remaining challenge is how to measure and credibly price the corresponding labor income risk exposure. Related to this, Lustig and Van Nieuwerburgh (2006) explore the required stochastic properties of the omitted components of wealth that are required to repair the model implications.

The use of aggregate nondurable consumption might also be too narrow. For this reason, many studies expand the definition of consumption and refine the preference assumptions when examining both the cross section and time series of asset returns. For example, Piazzesi, Schneider and Tuzel (2007) consider a separate role for housing, Yogo (2006) and Pakos (2006) examine the importance of consumer durables, and Uhlig (2006) considers leisure. Including these other components of consumption may also prove fruitful for our understanding of the wealth–consumption link. Further as emphasized by Uhlig (2006) these components are also germane to the evaluation of risk embedded in continuation values.

In this chapter we have been guilty of pushing the representative consumer model too hard. As an alternative to broadening the measure of wealth, we might focus on narrowing the definition of the marginal investor. Heaton and Lucas (2000) and others explore important aspects of investor heterogeneity, participation, market segmentation and limited risk sharing. Others, including Alvarez and Jermann (2000) and Lustig (2004) consider models in which there are important changes over time in the marginal investor participating in market. These changes induce an extra component to risk prices. All of these models provide alternative valuable frameworks for measurement. They do not, however, remove from consideration the modeling and measurement questions explored in this chapter.

Claims made in the empirical literature that intertemporal substitution can be inferred from the study of single asset returns such as Treasury bills or the risk free rate require qualification.[28] They ignore potentially important information that is often buried in the constant terms of log-linear estimation. We have seen how this additional information can rule out small values of the intertemporal substitution parameter (large values of ρ). The crude counterpart to this that abstracts from uncertainty can be seen by setting the subjective rate of discount to zero and comparing the growth rate of consumption to that of the average logarithm of returns. Excessively large values of our parameter ρ are inconsistent with the observed relation between means. While suggestive, this simple imitation of macro calibration is not formally correct in this context. As we have seen, the risk aversion parameter also comes into play. Separation can only be achieved as an approximation that abstracts from potentially important sample information.[29]

GMM inferences that explore shapes of the objective function through concentration or simulation are often the most revealing, even if they fail to achieve the simplified aims of Murray (2005). While the continuously-updated-GMM estimation has some advantages in terms of reliable inference, it can also reward parameter configurations that make the implied moment conditions hard to estimate. Thus naive use of such methods can lead to what turn out to be uninteresting successes. It is valuable to accompany such estimation with explorations of implied pricing errors or other assessments of potential misspecification.

Consumption-based models with long-run risk components pose interesting statistical challenges because they feature macroeconomic risk exposure over long horizons. Macroeconomic growth rate risk is reflected in continuation values, and continuation values contribute to risk prices defined both locally and in the long run. These prices along with cash-flow and return risk exposure determine the heterogeneity in asset prices. Investor risk preference is thus encoded in the predicted asset prices and expected returns. We have illustrated why this source of identifying information about investor risk preferences presents challenges for reliable measurement. Here we have illustrated this using VAR methods to assess such estimates. For more general specifications nonlinear solution methods and estimation methods will come into play.

The incorporation of more formal macroeconomics promises to aid our understanding of sources of long run risk. Work by Fisher (2007), Mulligan (2001) and others is suggestive of such links. Both use production-based macroeconomic models. Fisher focuses on long run potency of alternative sources of technology shocks. Mulligan (2001) considers consumption – physical return linkages as an alternative to the study of financial returns. Although stochastic volatility in consumption can potentially have long-run effects as well, this additional source of risk should ultimately have its source in shocks to technology and other economic fundamentals. Exploring these features in more fully specified models and focusing on long-run components hold promise for aiding our understanding of asset price heterogeneity.

[28] See Hansen and Singleton (1996), Campbell (2003) and Yogo (2004).

[29] Even in the power utility model with stochastic consumption, risk free rates are sometimes plausible with very large value of ρ as revealed by the volatility correction in a log-normal approximation.

Appendix A: Additional formulas for the Kreps–Porteus model

A.1. Discrete time

Recall that $v_t - c_t = U_v \cdot x_t + \mu_v$ where the formulas for U_v and μ_v are given in (21). Write

$$\left(v_t^1 - c_t\right)^2 = x_t' \Lambda x_t + 2\lambda' x_t + \ell.$$

We look for a solution for the derivative of the form

$$Dv_t^1 = -\left(\frac{1}{2} X_t' \Omega X_t + X_t \cdot \omega + \frac{\mathsf{w}}{2}\right)$$

where

$$\begin{aligned}
\Omega &= \frac{(1-\beta)}{\beta}\Lambda + \beta A' \Omega A, \\
\omega &= \frac{(1-\beta)}{\beta}\lambda + \beta(1-\gamma)A'\Omega B(H + B'U_v) + \beta A'\omega, \\
\mathsf{w} &= \frac{(1-\beta)}{\beta}\ell + \beta(1-\gamma)^2 (H + B'U_v)' B'\Omega B(H + B'U_v) \\
&\quad + 2\beta(1-\gamma)\omega' B(H + B'U_v) + \beta \operatorname{Tr}(B'\Omega B) + \beta \mathsf{w}.
\end{aligned} \tag{49}$$

The first equation in (49) is a Sylvester equation and is easily solved. Given Ω, the solution for ω is

$$\omega = (I - \beta A')^{-1}\left(\frac{1-\beta}{\beta}\lambda + \beta(1-\gamma)A'\Omega B(H + B'U_v)\right),$$

and given ω, the solution for w is obtained similarly by solving the third equation of (49).

Next we produce a formula for $Ds_{t+1,t}$ based on Equation (20). From our previous calculations

$$\begin{aligned}
&-(c_{t+1} - c_t) + \left[v_{t+1}^1 - \mathcal{Q}_t(v_{t+1}^1)\right] \\
&\quad = U_v' B w_{t+1} - G' x_t - \mu_c - \frac{1-\gamma}{2}\left|U_v' B + H'\right|^2.
\end{aligned}$$

Using our formulas for Dv_{t+1} for the distorted conditional expectation:

$$\begin{aligned}
&Dv_{t+1}^1 - E^*\left(Dv_{t+1}^1 | \mathcal{F}_t\right) \\
&\quad = -\frac{1}{2}\left(B w_{t+1}^*\right)' \Omega B w_{t+1}^* + \frac{1}{2}\operatorname{Tr}(B'\Omega B) \\
&\qquad - \left(B w_{t+1}^*\right)' \Omega\left[Ax_t + (1-\gamma)B(H + B'U_v)\right] - \omega' B w_{t+1}^*.
\end{aligned}$$

Substituting for w_{t+1}^* from the relation $w_{t+1} = w_{t+1}^* + (1-\gamma)[H + \beta B'(I - A'\beta)^{-1}G]$ we may implement formula (20) via,

$$Ds_{t+1,t}^1 = \frac{1}{2} w_{t+1}' \Theta_0 w_{t+1} + w_{t+1}' \Theta_1 x_t + \vartheta_0 + \vartheta_1 x_t + \vartheta_2 w_{t+1}$$

by constructing the coefficients Θ_0, Θ_1, ϑ_0, ϑ_1, ϑ_2.

A.2. Continuous time

In what follows, we derive the equations implied by (27) that can be used to compute the derivative of the value function in practice. Many readers may choose to skip this part.

To construct the solution, form the state vector

$$X_t = \begin{bmatrix} x_t \\ z_t \end{bmatrix}$$

and write composite state evolution (26) as

$$\mathrm{d}X_t = \widetilde{A}X_t\,\mathrm{d}t + \widetilde{F}\,\mathrm{d}t + \sqrt{z_t}\widetilde{B}_1\,\mathrm{d}W_t^* + \sqrt{z_t}\widetilde{B}_2\,\mathrm{d}\overline{W}_t^*,$$

and write

$$(U_v \cdot x + \overline{U}_v z + \mu_v)^2 = X'\Lambda X + 2\lambda' X + \ell.$$

Look for a derivative expressed as

$$Dv_t^1 = -\left(\frac{1}{2}X_t'\Omega X_t + X_t \cdot \omega + \frac{\mathsf{w}}{2}\right).$$

Substituting into Equation (27), Ω solves

$$-\delta\Lambda + \delta\Omega = \widetilde{A}'\Omega + \Omega\widetilde{A};$$

ω solves:

$$-\delta\lambda + \delta\omega = \Omega\widetilde{F} + \widetilde{A}'\omega + \begin{bmatrix} 0 \\ \frac{1}{2}\operatorname{Tr}(\Omega\widetilde{B}_1\widetilde{B}_1') + \frac{1}{2}\operatorname{Tr}(\Omega\widetilde{B}_2\widetilde{B}_2') \end{bmatrix};$$

and w solves:

$$-\delta\ell + \delta\mathsf{w} = 2\omega \cdot \widetilde{F}.$$

These three equations should be solved in sequence.

Given this solution we may compute the shock exposure vector for the derivative as follows:

$$\begin{bmatrix} D\sigma_{v,t}' \\ D\bar{\sigma}_{v,t} \end{bmatrix}' = -[\Omega X_t + \omega]'[\,\widetilde{B}_1 \quad \widetilde{B}_2\,] = -[\Omega X_t + \omega]'\begin{bmatrix} B & 0 \\ 0 & \overline{B} \end{bmatrix}.$$

Using these formulas, the risk prices are:

(i) $\mathrm{d}W_t$: $\sqrt{z_t}\rho H' + \sqrt{z_t}(\gamma - \rho)(B'U_v + H)' - \sqrt{z_t}(\rho - 1)(\gamma - 1)[\Omega X_t + \omega]'\tilde{B}_1$;

(ii) $\mathrm{d}\overline{W}_t$: $\sqrt{z_t}\rho\overline{H} + \sqrt{z_t}(\gamma - \rho)(\overline{B}\overline{U}_v + \overline{H}) - \sqrt{z_t}(\rho - 1)(\gamma - 1)[\Omega X_t + \omega]'\widetilde{B}_2$.

Appendix B: Bayesian confidence intervals

Consider the VAR:

$$A(L)y_t + C_0 + C_1 t = w_t$$

where y_{t+1} is d-dimensional. The matrix $A(0) = A_0$ is lower triangular. We base inferences on systems that can be estimated equation-by-equation. The w_t is assumed to be normal with mean zero and covariance matrix I. We follow Sims and Zha (1999) and Zha (1999) by considering a uniform prior on the coefficients and we follow Zha (1999) by exploiting the recursive structure of our models.

Write a typical equation as

$$\alpha z_t + \gamma \cdot x_t = v_t$$

where v_t is distributed as a standard normal, x_t is a vector of variables that are uncorrelated with v_t, but z_t is correlated with v_t. This equation can be transformed to a simple regression equation of z_t onto x_t with regression coefficients $\beta = -\frac{1}{\alpha}\gamma$ and regression error variance $\sigma^2 = \frac{1}{\alpha^2}$. Imposing a uniform prior over (α, γ) does not imply a uniform prior over the regression coefficients, however.

The piece of the likelihood for sample of T observations pertinent for this equation has the familiar form

$$\ell_T \propto |\alpha|^T \exp\left[-\sum_{t=1}^{T} \frac{(\alpha z_t + x_t \cdot \gamma)^2}{2}\right].$$

Consider first the posterior distribution of γ given α. Using familiar calculations e.g. see Box and Tiao (1973), it follows that

$$\gamma \sim \text{Normal}(-\alpha b_T, V_T)$$

where b_T is the least squares estimate obtained by regressing z_t onto x_t, and

$$V_T = \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1}.$$

The marginal posterior for α has a density that is proportional to

$$|\alpha|^T \exp\left(-\frac{\alpha^2 T s_T}{2}\right)$$

where s_T is the least squares residual variance

$$s_T = \frac{1}{T}\sum_{t=1}^{T}(z_t - x_t \cdot b_T)^2.$$

This is just a special case of a formula of Theorem 2 of Zha (1999).

It is convenient to use the distribution for $\upsilon = \alpha^2 T s_T$. By the change-of-variables formula the density for υ is proportional to

$$\upsilon^{\frac{T-1}{2}} \exp\left(-\frac{\upsilon}{2}\right),$$

which is the chi-square density with $T + 1$ degrees of freedom.

We simulate the joint posterior by first simulating υ using the chi-square distribution, then constructing α up to sign, and finally simulating γ conditioned on α according to a normal distribution.

Given the recursive nature of our model, we may follow Zha (1999) by building the joint posterior for all parameters across all equations as a corresponding product. This requires that we include the appropriate contemporary variables on the right-hand side of the equation to ensure that w_{t+1} has the identity as the covariance matrix. In effect we have divided the coefficients of the VAR into blocks that have independent posteriors given the data. We construct posterior confidence intervals for the objects that interest us a nonlinear functions of the VAR coefficients.

Appendix C: MCMC

The MCMC simulations follow a version of the standard Metropolis–Hastings algorithm [see Chernozhukov and Hong (2003)]. Let the parameter combination corresponding to the ith draw be $b^{(i)} = [\delta^{(i)}, \gamma^{(i)}]$ (since we hold ρ constant in these simulations, we omit reference to it here). Then

1. draw $b^{(0)}$ from the prior distribution (uniform on A);
2. draw ζ from the conditional distribution $q(\zeta|b^{(i)})$;
3. with probability $\inf\left(\frac{\exp(-\psi_T(b^{(i+1)}))q(b^{(i)}|\zeta)}{\exp(-\psi_T(b^{(i)}))q(\zeta|b^{(i)})}, 1\right)$ update $b^{(i+1)} = \zeta$; otherwise keep $b^{(i+1)} = b^{(i)}$.

A typical choice of transition density is Gaussian, which results in a Markov chain that is a random walk. We are interested in constraining the parameter space to a compact set. Therefore an adjustment needs to be made for truncating the distribution. Specifically, let ϕ be the bivariate normal density centered around zero with cdf Φ. Then

$$q(x|y) = \frac{\phi(x-y)}{\Pr(x \in A)}, \quad \text{where } x = y + z,\ z \sim \Phi,$$

which can be computed straightforwardly. In simulations, the truncation is accomplished by discarding the values of ζ that fall outside of A. A choice needs to be made regarding the dispersion of ϕ. Too large a variance would generate too many truncations and thus result in slow convergence while too low a value would produce a very slowly-moving random walk that might fail to visit substantial regions of the parameter space and also lead to slow convergence. We set the standard deviations of ϕ for both parameters equal to their respective ranges divided by 50. The reported results are based on simulations with 1,000,000 draws.

Appendix D: Data description

Data: population is from NIPA Table 2.1. Risk-free rate is the 3-month Treasury Bill rate obtained from CRSP Fama Risk Free Rates files.

Book-to-market portfolios: Returns to value weighted portfolios of stocks listed on NASDAQ, AMEX and NYSE. Stocks sorted by book-to-market value of equity. Construction of these portfolio returns is detailed in Hansen, Heaton and Li (2005).

Consumption: Aggregate US consumption of nondurables and services as reported in the National Income and Product Accounts of the United States. Seasonally adjusted and converted to real units using the implicit price deflators for nondurables and services. Quarterly from 1947 to 2006.

Corporate earnings: "Corporate profits with IVA and CCAdj" from the National Income and Product Accounts of the United States. Quarterly, seasonally adjusted from 1947 to 2005.

Dividends: Constructed from the portfolio returns "with" and "without" dividends. Seasonality removed by taking a moving average. Construction of this series is detailed in Hansen, Heaton and Li (2005).

Market return: Value weighted return to holding stocks listed on NASDAQ, AMEX and NYSE. Constructed from CRSP data base. Quarterly from 1947 to 2006.

Population: US civilian noninstitutionalized population 1947 to 2005.

Price deflator: Implicit price deflator for nondurables and services. Quarterly from 1947 to 2005.

Risk free rate: Three-month Treasury Bill return from CRSP. Quarterly from 1947 to 2006.

Wages and salaries: Wages and salary disbursement from the National Income and Product Accounts of the United States. Seasonally adjusted and converted to real units using the implicity price deflators for nondurables and services. Quarterly from 1947 to 2005.

Wealth: Total financial assets of the United States personal sector less Total liabilities as reported in table L.10 of the Flow of Funds Accounts of the United States. Quarterly from 1952 to 2005.

References

Abel, A. (1990). "Asset prices under habit formation and catching up with the Jones". American Economic Review 80, 38–42.

Alvarez, F., Jermann, U.J. (2000). "Efficiency, equilibrium and asset pricing with risk of default". Econometrica 68 (4), 775–797.

Anderson, T.W., Rubin, H. (1949). "Estimation of the parameters of a single equation in a complete system of stochastic equations". Annals of Mathematical Statistics 20, 46–63.

Backus, D.K., Routledge, B.R., Zin, S.E. (2004). "Exotic preferences for macroeconomics". In: Gertler, M., Rogoff, K. (Eds.), NBER Macroeconomics Annual 2004.

Bansal, R., Lehmann, B.N. (1997). "Growth optimal portfolio restrictions on asset pricing models". Macroeconomic Dynamics 1, 333–354.
Bansal, R., Yaron, A. (2004). "Risks for the long run: A potential resolution of asset pricing puzzles". Journal of Finance 59 (4), 1481–1509.
Bansal, R., Dittmar, R.F., Lundblad, C.T. (2005). "Consumption, dividends, and the cross-section of equity returns". Journal of Finance 60 (4), 1639–1672.
Bekaert, G., Hodrick, R.J., Marshall, D.A. (1997). "The implications of first-order risk aversion for asset market premiums". Journal of Monetary Economics 40, 3–39.
Blanchard, O.J., Quah, D. (1989). "The dynamic effects of aggregate demand and supply disturbances". American Economic Review 79, 655–673.
Box, G.E.P., Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Addison–Wesley, Reading, MA.
Breeden, D. (1979). "An intertemporal asset pricing model with stochastic consumption and investment opportunities". Journal of Financial Economics 7, 265–296.
Campbell, J.Y. (1996). "Understanding risk and return: A consumption-based explanation of aggregate stock market behavior". Journal of Political Economy 104, 298–345.
Campbell, J.Y. (2003). "Consumption-based asset pricing". In: Harris, M., Constantinides, G.M., Stulz, R.M. (Eds.), Handbook of the Economics of Finance. Elsevier.
Campbell, J.Y., Cochrane, J.H. (1999). "By force of habit". Journal of Political Economy 107, 205–251.
Campbell, J.Y., Shiller, R.J. (1988a). "The dividend–price ratio and expectations of future dividends and discount factors". Review of Financial Studies 1, 195–227.
Campbell, J.Y., Shiller, R.J. (1988b). "Stock prices, earnings, and expected dividends". Journal of Finance 43, 661–676.
Campbell, J.Y., Viceira, L.M. (2002). Strategic Asset Allocation: Portfolio Choice for Long-Term Investors. Oxford University Press, Oxford, UK.
Campbell, J.Y., Vuolteenaho, T. (2004). "Bad beta, good beta". American Economic Review 94, 1249–1275.
Chen, X., Ludvigson, S.C. (2004). "Land of addicts? An empirical investigation of habit-based asset pricing behavior". NBER Working Paper 10503. New York University.
Chernozhukov, V., Hong, H. (2003). "An MCMC approach to classical estimation". Journal of Econometrics 115, 293–346.
Cochrane, J.H. (1992). "Explaining the variance of price–dividend ratios". Review of Financial Studies 5, 243–280.
Cochrane, J.H., Hansen, L.P. (1992). "Asset pricing explorations for macroeconomics". In: NBER Macroeconomics Annual.
Constantinides, G.M. (1990). "Habit formation: A resolution of the equity premium puzzle". Journal of Political Economy 98, 519–537.
Constantinides, G., Duffie, D. (1996). "Asset pricing with heterogeneous consumers". Journal of Political Economy 104, 219–240.
Duffie, D., Epstein, L.G. (1992a). "Asset pricing with stochastic differential utility". Review of Financial Studies 5 (3), 411–436.
Duffie, D., Epstein, L.G. (1992b). "Stochastic differential utility". Econometrica 60 (2), 353–394.
Epstein, L., Schneider, M. (2003). "Recursive multiple priors". Journal of Economic Theory 113 (1), 1–31.
Epstein, L., Zin, S. (1989a). "Substitution, risk aversion and the temporal behavior of consumption and asset returns: A theoretical framework". Econometrica 57, 937–969.
Epstein, L., Zin, S. (1989b). "Substitution, risk aversion and the temporal behavior of stock returns: An empirical investigation". Journal of Political Economy 99, 263–286.
Epstein, L.G., Zin, S. (1990). "First-order risk aversion and the equity premium puzzle". Journal of Monetary Economics 26, 387–407.
Fama, E., French, K. (1992). "The cross-section of expected stock returns". Journal of Finance 47, 427–465.
Fisher, J. (2007). "The dynamic effects of neutral and investment-specific technology shocks". Journal of Political Economy 115 (1), 141–168.
Gallant, A.R., Hansen, L.P., Tauchen, G. (1990). "Using conditional moments of asset payoffs to infer the volatility of intertemporal marginal rates of substitution". Journal of Econometrics 45, 141–179.

Garcia, R., Renault, E., Semenov, A. (2006). "Disentangling risk aversion and intertemporal substitution". Finance Research Letters 3, 181–193.
Grossman, S.J., Shiller, R.J. (1981). "The determinants and variability of stock market prices". American Economic Review 71, 222–227.
Gul, F. (1991). "A theory of disappointment aversion". Econometrica 59, 667–686.
Hall, R.E. (1988). "Intertemporal substitution and consumption". Journal of Political Economy 96, 339–357.
Hansen, L.P. (1982). "Large sample properties of generalized method of moments estimators". Econometrica 50, 1029–1054.
Hansen, L.P. (2004). "Comment on exotic preferences for macroeconomics". In: Gertler, M., Rogoff, K. (Eds.), NBER Macroeconomics Annual 2004.
Hansen, L.P. (2006). "Modeling the long run: Valuation in dynamic stochastic economies". Fisher–Schultz lecture at the European meetings of the econometric society.
Hansen, L.P., Jagannathan, R. (1991). "Restrictions on intertemporal marginal rates of substitution implied by asset returns". Journal of Political Economy 99, 225–262.
Hansen, L.P., Jagannathan, R. (1997). "Assessing specification errors in stochastic discount factor models". Journal of Finance 52, 557–590.
Hansen, L.P., Richard, S.J. (1987). "The role of conditioning information in deducing testable implications of asset pricing models". Econometrica 55 (3), 587–613.
Hansen, L.P., Sargent, T. (1995). "Discounted linear exponential quadratic Gaussian control". IEEE Transactions on Automatic Control 40 (5), 968–971.
Hansen, L.P., Scheinkman, J. (2006). "Long run risk: An operator approach". Working paper. University of Chicago and Princeton University.
Hansen, L.P., Singleton, K. (1982). "Generalized instrumental variables estimation of nonlinear rational expectations models". Econometrica 50, 1269–1299.
Hansen, L.P., Singleton, K. (1983). "Stochastic consumption, risk aversion, and the temporal behavior of asset returns". Journal of Political Economy 91, 249–265.
Hansen, L.P., Singleton, K. (1996). "Efficient estimation of linear asset pricing models". Journal of Business and Economic Statistics 14, 53–68.
Hansen, L.P., Roberds, W., Sargent, T.J. (1991). "Observable implications of present-value-budget balance". In: Hansen, L.P., Sargent, T.J. (Eds.), Rational Expectations Econometrics. Westview Press.
Hansen, L.P., Heaton, J., Luttmer, E. (1995). "Econometric evaluation of asset pricing models". Review of Financial Studies 8, 237–274.
Hansen, L.P., Heaton, J., Yaron, A. (1996). "Finite sample properties of some alternative GMM estimators". Journal of Business and Economic Statistics 14, 262–280.
Hansen, L.P., Heaton, J., Li, N. (2005). "Consumption strikes back?: Measuring long-run risk". NBER Working Paper 11476. National Bureau of Economic Research.
Hansen, L.P., Sargent, T.J., Turmuhambetova, G.A., Williams, N. (2006). "Robust control and model misspecification". Journal Economic Theory 128 (1), 45–90.
Harrison, J., Kreps, D. (1979). "Martingales and arbitrage in multiperiod security markets". Journal of Mathematical Economics 20, 381–408.
Heaton, J. (1995). "An empirical investigation of asset pricing with temporally dependent preferences". Econometrica 63, 681–717.
Heaton, J., Lucas, D. (2000). "Stock prices and fundamentals". In: NBER Macroeconomics Annual, 1999.
Jacobson, D.H. (1973). "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games". IEEE Transactions for Automatic Control AC-18, 1124–1131.
Jagannathan, R., Wang, Z. (1996). "The conditional CAPM and the cross section of stock returns". Journal of Finance 54, 1325–1361.
Kleibergen, F. (2005). "Testing parameters in GMM without assuming that they are identified". Econometrica 73, 1103–1123.
Kogan, L., Uppal, R. (2001). Risk Aversion and Optimal Portfolio Policies in Partial and General Equilibrium Economies. Sloan School of Management and London Business School.

Kreps, D.M., Porteus, E.L. (1978). "Temporal resolution of uncertainty and dynamic choice". Econometrica 46, 185–200.

Lettau, M., Ludvigson, S. (2001). "Consumption, aggregate wealth and expected stock returns". Journal of Finance LVI, 815–849.

Lettau, M., Ludvigson, S. (2003). "Measuring and modeling variation in the risk-return tradeoff". In: Ait-Sahalia, Y., Hansen, L.P. (Eds.) Handbook of Financial Econometrics. Elsevier. In press.

Lettau, M., Ludvigson, S. (2004). "Understanding trend and cycle in asset values: Reevaluating the wealth effect on consumption". American Economic Review 94, 276–299.

Lettau, M., Wachter, J. (2007). "Why is long-horizon equity less risky? A duration-based explanation of the value premium". Journal of Finance 62, 55–92.

Lucas, R.E. (1978). "Asset prices in an exchange economy". Econometrica 46, 1429–1445.

Lustig, H. (2004). "The market price of aggregate risk and the wealth distribution". Working Paper. UCLA.

Lustig, H., Van Nieuwerburgh, S. (2006). "The returns on human capital: Good news on wall street is bad news on main street". Review of Financial Studies. In press.

Luttmer, E.G.J. (1996). "Asset pricing with frictions". Econometrica 64, 1439–1467.

Maenhout, P.J. (2004). "Robust portfolio rules and asset pricing". Review of Financial Studies 17, 951–983.

Malloy, C., Moskowitz, T., Vissing-Jorgensen, A. (2005). "Long-run stockholder consumption and asset returns". Working Paper.

Mehra, R., Prescott, E. (1985). "The equity premium: A puzzle". Journal of Monetary Economics 15, 145–161.

Merton, R.C. (1973). "An intertemporal capital asset pricing model". Econometrica 41 (5), 867–887.

Mulligan, C. (2001). Capital Interest, and Aggregate Intertemporal Substitution. University of Chicago.

Murray, M.P. (2005). "The bad, the weak, and the ugly: Avoiding the pitfalls of instrumental variables estimation". Working Paper.

Newey, W.K., West, K. (1987). "A simple positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix". Econometrica 55, 703–708.

Novales, A. (1990). "Solving nonlinear rational expectations models: A stochastic equilibrium model of interest rates". Econometrica 58, 93–111.

Pakos, M. (2006). "Asset pricing with durable goods and non-homothetic preferences". Manuscript. Carnegie Mellon University.

Parker, J.A., Julliard, C. (2005). "Consumption risk and the cross section of expected returns". Journal of Political Economy 113 (1), 185–222.

Petersen, I.R., James, M.R., Dupuis, P. (2000). "Minimax optimal control of stochastic uncertain systems with relative entropy constraints". IEEE Transactions on Automatic Control 45, 398–412.

Piazzesi, M., Schneider, M., Tuzel, S. (2007). "Housing, consumption, and asset pricing". Journal of Financial Economics 83 (3), 531–569.

Restoy, F., Weil, P. (1998). "Approximate equilibrium asset prices". NBER Working Paper 6611.

Roussanov, N. (2005). "Composition of wealth, conditioning information, and the cross-section of stock returns". Working paper.

Routledge, B.R., Zin, S.E. (2003). "Generalized disappointment aversion and asset prices". NBER Working Paper 10107.

Sargan, J.D. (1958). "Estimation of economic relations using instrumental variables". Econometrica 26, 393–415.

Shiller, R. (1981). "Do stock prices move too much to be justified by subsequent changes in dividends?" American Economic Review 71 (3), 421–436.

Shiller, R.J. (1982). "Consumption, asset markets and macroeconomic fluctuations". Carnegie–Rochester conference series on public policy 17, 203–238.

Sims, C.A., Zha, T. (1999). "Error bands for impulse responses". Econometrica 67 (5), 1113–1156.

Skiadas, C. (2003). "Robust control and recursive utility". Finance and Stochastics 7, 475–489.

Snow, K.N. (1991). "Diagnosing asset pricing models using the distribution of asset returns". Journal of Finance 46, 955–984.

Stock, J.H., Wright, J.H. (2000). "GMM with weak identification". Econometrica 68, 1055–1096.
Stutzer, M. (1996). "A simple nonparametric approach to derivative security valuation". Journal of Finance 51, 1633–1652.
Tallarini, T. (1998). "Risk sensitive real business cycles". Journal of Monetary Economics 45 (3), 507–532.
Uhlig, H. (2006). "Asset pricing with Epstein–Zin preferences". Working Paper. Humboldt University, Berlin.
Vuolteenaho, T. (2002). "What drives firm-level stock returns?" Journal of Finance 57, 233–264.
Wang, K.Q. (2003). "Asset pricing with conditioning information: A new test". Journal of Finance 58, 161–196.
Weil, P. (1989). "The equity premium puzzle and the risk-free rate puzzle". Journal of Monetary Economics 24, 401–421.
Whittle, P. (1990). Risk-Sensitive Optimal Control. John Wiley & Sons, New York.
Yogo, M. (2004). "Estimating the elasticity of intertemporal substitution when instruments are weak". Review of Economics and Statistics 86 (3), 797–810.
Yogo, M. (2006). "A consumption-based explanation of expected stock returns". Journal of Finance 61 (2), 539–580.
Zha, T. (1999). "Block recursion and structural vector autoregression". Journal of Econometrics 90, 291–316.

Chapter 62

A PRACTITIONER'S APPROACH TO ESTIMATING INTERTEMPORAL RELATIONSHIPS USING LONGITUDINAL DATA: LESSONS FROM APPLICATIONS IN WAGE DYNAMICS*

THOMAS MACURDY[1]

The Hoover Institution, Stanford University, Stanford, CA 94305, USA
e-mail: tmac@stanford.edu

Contents

* The author gratefully acknowledges research support from NIH grants HD32055-02 and 5P01AG005842-17. This chapter benefitted greatly from participants' comments at the Chicago and London Handbook conferences and from expert research assistance provided by Frank McIntyre and James Wishart Pearce.

[1] Professor, Department of Economics, and Senior Fellow, The Hoover Institution, Stanford University.

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06062-X

Abstract

This chapter presents a unified set of estimation methods for fitting a rich array of models describing dynamic relationships within a longitudinal data setting. The discussion surveys approaches for characterizing the micro dynamics of continuous dependent variables both over time and across individuals, focusing on two flexible sets of empirical specifications: dynamic simultaneous equations models incorporating error-components structures, and autoregressive quantile models. The chapter is motivated by the principle that, whenever possible, estimation methods should rely on routines available in familiar software packages to make them accessible to a wide range of practitioners. Conventional method-of-moments procedures offer a general apparatus for estimating parameters of panel-data specifications, though one must introduce a series of modifications to overcome challenges arising from: (1) use of unbalanced data structures, (2) weighting to account for stratified sampling inherent in survey longitudinal data, (3) incorporation of predetermined variables in estimation, and (4) computational complexities confronted when estimating large systems of equations with intricate intertemporal restrictions. To allow researchers to separate the estimation of longitudinal time-series specifications into manageable pieces, the discussion describes multi-step approaches that estimate subsets of parameters appearing in a single model component (such as the autoregressive or moving-average structure of the error process) without having to estimate all parameters of the entire model jointly. Such procedures offer a powerful set of diagnostic tools for narrowing model choices and for selecting among specifications that fit the underlying data. To illustrate all of the econometric methods outlined in this chapter, the analysis presents a set of empirical applications summarizing the dynamic properties of hourly wages for adult men using data from the Panel Study of Income Dynamics.

Keywords

earnings dynamics, longitudinal data, dynamic simultaneous equations, dynamic quantile regressions, error structure, nonlinear simultaneous equations, method of moments, optimal instruments, sample weighting, stratified sample, unbalanced data, multi-step estimation, autoregressive, ARMA, times series

JEL classification: C10, C13, C15, C21, C22, C23, C31, C32, C33, C51, C52, C53, D90, J31, J60

1. Introduction

Few topics in empirical economics have received as much attention as the dynamic properties of wages and earnings. The questions asked in this work include: To what extent does the dispersion of an individual's earnings grow over time? Is this dispersion shared by other individuals and groups? Does this dispersion give rise to a shift in rankings of individuals within groups? Comprehensive answers to such questions require knowledge of two factors that jointly determine the dynamic properties of an individual's earnings: market-wide trends governing the evolution of cross-sectional distributions over time, and forces underlying an individual's mobility within distributions. Empirical analyses exploiting micro-longitudinal data constitute a prominent approach for acquiring this knowledge, an approach that relies on a rich array of econometric specifications to uncover the intertemporal relationships linking an economic agent's variables over both short and long time horizons. This chapter overviews the panel data models and estimation methods found in the literature on earnings and income dynamics. While it motivates the discussion by drawing upon the questions and analyses found in this extensive literature, one can readily apply the empirical methods covered here to characterize the intertemporal properties of a wide range of variables available in longitudinal settings.

The discussion surveys methods for estimating specifications designed to characterize the dynamic properties of continuous dependent variables in panel data settings, variables whose stochastic fluctuations follow patterns consistent with sophisticated forms of time-series and error-component models. In addition to considering flexible intertemporal specifications for error structures, the analysis admits nonlinear structural equations as a vehicle for relating measured variables both contemporaneously and over time. This chapter not only covers such specifications that provide a summary of the autocorrelation patterns of variables, which link the first and second moments of variables over time, but it also explores procedures for fitting quantiles to describe dynamic relationships. Panel data offer multiple observations on individuals over several periods, often supplying only short and noncontiguous time series for members of a large cross section of individuals. This feature of longitudinal data means that not all of the specifications and estimation procedures applicable in conventional time series analysis carry over to panel data, and, alternatively, many of the problems and options relevant in analyzing longitudinal data are not found in standard time series studies.

Familiar "method of moments" (MM) procedures provide a general apparatus for estimating parameters of panel data specifications, but one encounters a variety of challenges in implementing such procedures in longitudinal data settings. One issue, easily overcome by drawing upon findings in the literature, concerns how to exploit predetermined variables – quantities which are endogenous in some equations but not in others – as instrumental variables in estimation. More demanding challenges involve computational complexities confronted when estimating large systems of equations with intricate nonlinearities, circumstances that often come about in panel data applications, especially when one incorporates empirical specifications to estimate dynamic struc-

tures describing error processes. Still more formidable challenges concern how to use weights to account for the stratified samples that are a part of all longitudinal surveys, and how to carry out estimation with unbalanced samples – samples that have an uneven number of, and possibly different, time series observations across individuals. Longitudinal surveys supply a variety of weights for use in the calculation of statistics to compensate for nonrandom sampling, and the question arises as to which weights one should use in MM procedures when estimating dynamic relationships. Use of unbalanced datasets in MM procedures to avoid discarding data typically leads to the reporting of invalid standard errors and test statistics by conventional MM formulas.

To surmount the computational challenges one can encounter with implementation of MM procedures, this chapter lays out several options enabling practitioners to estimate sophisticated longitudinal data specifications using standard routines available in familiar software packages. Linear/nonlinear 3SLS procedures rely on convenient computational formulas for large systems of equations. While 3SLS routines do not allow for the inclusion of predetermined variables as instrumental variables, the subsequent discussion demonstrates how to modify a conventional 3SLS system to exploit predetermined variables fully in estimation with a minimal amount of extra programming and computational burden. In addition, to allow researchers to separate the estimation of longitudinal time-series specifications into manageable pieces, the discussion describes multi-step approaches. When carrying out a step, a researcher can focus on estimating only the subset of parameters appearing in a single model component (such as the AR or MA structure of the error process) without having to estimate all parameters of the entire model jointly. Such procedures offer a powerful set of diagnostic tools for narrowing model choices and for selecting among specifications that fit the underlying data.

Regarding other challenges, this chapter demonstrates how to incorporate weights in MM procedures to compensate for the nonrandom sampling frames inherent in longitudinal surveys – thus avoiding naive MM implementations that produce inconsistent estimates and/or test statistics. It also describes how to construct weights to enable use of unbalanced data structures. In addition to describing estimation of dynamic simultaneous equations that relate the moments and autocorrelation patterns of earnings over time, the analysis also outlines how this apparatus can be applied to estimate specifications characterizing the autoregressive properties of the quantiles of earnings.

To illustrate implementation of the econometric methods outlined in this chapter, the exposition relies on a unified set of empirical applications rather than attempting to cite examples from the existing literature; the current body of studies does not offer sufficient overlap or cover all issues necessary to exemplify approaches. All empirical illustrations presented here utilize a common dataset on men's wages drawn from the Panel Study of Income Dynamics (PSID). The discussion presents ideas in a way useful to practitioners who wish to specify and estimate models capable of addressing their empirical questions, not to readers desiring a knowledge of the rigorous theoretical underpinnings of econometric results or a comprehensive documentation of studies in

the field. While the chapter draws heavily on existing results in the literature and does not claim originality of the concepts outlined here, it does not attempt to attribute each development to specific authors and omits many references. Instead, as concepts are introduced, it directs readers to a variety of other surveys, especially to other chapters in this and other Handbooks, that offer a wealth of citations and references along with alternative presentations of the material.

This chapter does not address a variety of interesting topics important in analyses of longitudinal data and aggregate trends. First, this chapter focuses on estimation approaches applicable for continuous dependent variables, and not for dependent variables that are discrete, censored, or truncated. For discrete variables in a longitudinal setting, popular estimation approaches include duration and competing risk models, topics comprehensively covered in *Handbook of Econometrics* chapters by Heckman and Singer (1986) and van den Berg (2001). Handbook chapters by McFadden (1984), Hajivassiliou and Ruud (1994), and Arellano and Honoré (2001) describe other valuable approaches for estimating dynamic relationships involving discrete, as well as censored and truncated, variables. Second, this chapter interprets specifications of market-wide trends as time effects that are common across population segments, and estimates these effects as parameters. Therefore, the analysis does not consider the more elaborate specifications for aggregate trends that incorporate sophisticated forms of stochastic components of the sort entertained in Handbook chapters surveying time-series techniques by Granger and Watson (1984), Hendry, Pagan and Sargan (1984), Geweke (1984), Wooldridge (1994), Stock (1994), Watson (1994), Teräsvirta, Tjøstheim and Granger (1994) and Hamilton (1994); nor does this chapter survey the empirical methods found in the extensive literature documenting the market-wide trends in wage growth and earnings inequality that have occurred during the past three decades, a topic covered in the *Handbook of Labor Economics* chapter by Katz and Autor (1999). Third, this chapter restricts attention to classical estimation methods applicable for parametric specifications; consequently, it does not cover the burgeoning literature on nonparametric and semi-parametric estimation methods, nor does it address the use of Bayesian procedures. Handbook chapters by Härdle and Linton (1994), Powell (1994), Geweke and Keane (2001) and Abbring, Heckman and Vytlacil (2007) provide insightful overviews of these alternative estimation methods.

Six sections make up the core of this chapter. Section 2 surveys the wide variety of empirical specifications applied in the longitudinal data literature to characterize the dynamic properties of wages and earnings, considering specifications for both error structures and distributed lag relationships linking measured variables. Section 3 briefly covers the key asymptotic results underlying MM estimation and testing procedures, along with the challenges encountered in implementing these methods in panel data settings. Section 4 presents several approaches for simplifying the estimation problem confronted when fitting sophisticated longitudinal specifications, with the focus on subdividing the overall estimation problem into a series of manageable steps. Section 5 outlines how to adapt the empirical apparatus covered in the earlier sections to provide for estimation of autoregressive specifications for the quantiles of variables. Section 6

describes how to integrate the weights provided by longitudinal surveys into the estimation of panel data models, and it goes on to develop a modified weighting-type procedure enabling one to use unbalanced data structures to estimate dynamic specifications. To illustrate application of the modeling and theoretical concepts covered in this chapter, Section 7 presents a series of empirical examples designed to explore the dynamic properties of the hourly wages of men using data from the PSID for the years 1980–1991. The purpose of this empirical analysis is not only to enhance accessibility to practitioners, but also to offer some insights into the sensitivities of findings induced by relying on alternative methods. As a conclusion to the overall chapter, Section 8 offers an abbreviated summary and concluding remarks.

2. Empirical specifications describing panel data dynamics

The empirical literature characterizing wage and earnings dynamics in panel data settings exploits a wide variety of specifications. Modeling the intertemporal properties of continuously-distributed variables using longitudinal surveys involves distinguishing two sources of variation in data: aggregate dynamics determining how cross-sectional distributions evolve over time, and micro dynamics describing the evolution of individual agents' relative positions within cross-sectional distributions from period to period. This chapter reviews empirical approaches focused on characterizing the micro aspects of dynamics.

Two components make up panel data specifications designed to capture the underlying sources of micro dynamics experienced by individuals. The first relates to measured variables, be they endogenous, predetermined or exogenous quantities. These specifications may be nonlinear in both parameters and variables; they often incorporate distributed lag relationships. The second model component describes the stochastic properties of error terms appearing in structural equations. These properties reflect features of the time series processes generating individual-specific errors. One finds elaborate representations for these time-series models in longitudinal data analyses, including sophisticated integration of nonstationary ARMA specifications and error-component models comprised of time-varying combinations of individual-specific factors.

This section begins with a broad characterization of the empirical specifications whose estimation occupies the remainder of this chapter. After identifying how such specifications account for the underlying aggregate and micro aspects of dynamics, the discussion reviews the empirical parameterizations applied to model complex autocorrelation structures linking both measured variables and unobserved errors over time. In addition to exploring parameterizations that describe the intertemporal pattern of the moments of variables, this section ends with a discussion of empirical parameterizations aimed at modeling the dynamic properties of dependent variables through the evolution of quantiles over time.

2.1. General characterization of empirical specifications

The panel data models discussed in this survey belong to parameterizations of the following nonlinear simultaneous equation:

$$f_{ti} = f_{ti}(Y_{ti}, Z_{ti}, X_{ti}, \gamma) = U_{ti}. \tag{2.1}$$

The function f_{ti} possesses a known form, but the parameter vector γ is unknown and must be estimated. The data vectors Y_{ti}, Z_{ti} and X_{ti} have the structure

$$Y_{ti} = \begin{pmatrix} y_{ti} \\ \vdots \\ y_{(t-k_y)i} \end{pmatrix}, \qquad Z_{ti} = \begin{pmatrix} z_{ti} \\ \vdots \\ z_{(t-k_z)i} \end{pmatrix}, \qquad X_{ti} = \begin{pmatrix} x_{ti} \\ \vdots \\ x_{(t-k_x)i} \end{pmatrix}, \tag{2.2}$$

with observations available for "agent" or "individual" i in time period t. The models considered here assume that a panel dataset supplies T time series observations for each of N cross sectional observations on individuals. The y_{ti}'s in (2.2) represent current and lagged endogenous variables; the z_{ti}'s comprise additional sets of endogenous and predetermined quantities; and the x_{ti}'s constitute exogenous variables. The quantities f_{ti}, y_{ti}, z_{ti} and x_{ti} may all be interpreted as column vectors, but much of the discussion treats them as scalars to simplify the exposition.

The error term U_{ti} in (2.1) follows a generalized ARMA(p, q) process given by

$$U_{ti} = -\sum_{j=1}^{p} a_{jt} U_{(t-j)i} + \sum_{j=0}^{q} m_{jt} \varepsilon_{(t-j)i},$$

where the ε_{ti}'s constitute mean-zero disturbances that are independently distributed over both time and individuals, and the a_{jt}'s and m_{jt}'s are parameters with $a_{0t} = m_{0t} = 1$. A compact representation of this equation is

$$a_t(L) U_{ti} = m_t(L) \varepsilon_{ti}, \tag{2.3}$$

where $a_t(L) \equiv \sum_{j=0}^{p} a_{jt} L^j$ and $m_t(L) \equiv \sum_{j=0}^{q} m_{jt} L^j$ are lag polynomials of orders p and q respectively.[1] In many longitudinal data specifications, the coefficients of $a_t(L)$ and $m_t(L)$ are time invariant. The error terms ε_{ti} are independently distributed over time and individuals, with a variance–covariance structure given by

$$E(\varepsilon_{ti} \varepsilon_{ks}) = \begin{cases} \Sigma_{ti} & \text{if } i = s \text{ and } t = k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

When $\Sigma_{ti} = \Sigma_{ki} \equiv \Sigma_i$ for all t and k, the literature designates the ε_{ti}'s as white noise, for they satisfy a weak stationarity property (i.e., have constant variances over time). When $\Sigma_{ti} = \Sigma_{tj} \equiv \Sigma_t$ for all i and j, the disturbances ε_{ti} are homoscedastic

[1] The roots of the polynomial $m_t(L)$ are assumed to lie on or outside the unit circle. This restriction is the usual one imposed in the time series literature to guarantee identification of the coefficients of $m_t(L)$.

across individuals. The subsequent discussion covers estimation procedures allowing for the parameters Σ_{ti} to be constant over time and/or across agents. Regardless of the specification of Σ_{ti}, the following exposition refers to the ε_{ti}'s as white noise.

2.2. Sources of dynamics

Modeling the dynamic properties of variables for individuals requires distinguishing two sources of variation: components reflecting shared time effects that jointly displace measures for entire groups, and individual-specific sources of variation. In practice, there can be considerable discretion in attributing fluctuations to these different sources, making this conceptually-simple task quite difficult. Often this task is accomplished without researchers explicitly realizing that they have done so.

To fix ideas, consider the simple linear regression specification

$$y_{ti} = \pi y_{(t-1)i} + \beta_1 x_{ti} + \beta_2 x_{(t-1)i} + \tau_t + \nu_{ti}, \quad t = 1, \ldots, T, \ i = 1, \ldots, N, \tag{2.5}$$

where τ_t constitutes a time effect common to all individuals i in period t, and the error terms ν_{ti} are distributed independently of all time components making up the vector of $\tau' \equiv (\tau_1 \ldots \tau_T)$. According to this relation, time effects influence the dynamics of y_{ti}'s only by shifting the means of the cross-sectional distributions from one period to the next. If, in addition, the variance of the ν_{ti} systematically grows or declines over time, then one might replace (2.5) by

$$y_{ti} = \pi y_{(t-1)i} + \beta_1 x_{ti} + \beta_2 x_{(t-1)i} + \tau_{1t} + \tau_{2t} \nu_{ti}^*, \tag{2.6}$$

where τ_{1t} and τ_{2t} are distinct time components realized in period t, and now ν_{ti}^* is distributed independently of all the τ_{1t}'s and τ_{2t}'s. When there are multiple time effects operative in period t, the subsequent analysis interprets τ_t as a vector incorporating these effects, and the vector τ as including all time components τ_t.[2]

Inspection of (2.5) reveals that the overall dynamic properties of y_{ti} depend on four sources: (1) the stochastic behavior of the time components τ_t; (2) the parameters determining the distributed lag relationships involving past y_{ti}'s and x_{ti}'s; (3) the intertemporal properties of the x_{ti}'s; and (4) the stochastic process of the errors ν_{ti}.

2.2.1. Aggregate dynamics

Many studies focus on understanding the forces underlying trends in the economy or in a market, rather than how individuals sort themselves around these trends from one period to the next. These forces determine the evolution of cross-sectional distributions

[2] Analyses may further specify that time effects differ across groups of individuals, in which one might further substitute $\tau_{g_1 t}$ and $\tau_{g_2 t}$ for τ_{1t} and τ_{2t} in (2.6) where the subscript "g" distinguishes particular groups. In this case, the notation τ_t would be a vector incorporating the elements $\tau_{g_1 t}$ and $\tau_{g_2 t}$ for all groups.

over time. This exercise requires knowledge of the time patterns followed by the τ_t's, for movements in the τ_t's determine how cross-sectional distributions shift over time.

In a micro empirical analysis, one can treat the τ_t's as fixed or random effects. The majority of micro studies estimate the elements of τ as parameters, thus implicitly interpreting the τ_t's as fixed effects. Another popular approach treats the time effects τ_t as deterministic functions of exogenous variables, with year and age variables introduced to capture underlying trends. Such analyses often abandon attempts to learn much about the intertemporal process determining time effects and merely plot the estimated values of $\tau_1, \ldots, \tau_T$ against time.

When studies interpret the τ_t's as random time effects, their purpose is to estimate relationships characterizing the stochastic properties of these time components. Typically, micro analyses interpret the time effects as being independently distributed over time. In contrast, macroeconomic analyses introduce sophisticated relationships to model the dynamic properties of these economy-wide effects, exploiting rich specifications based on ARMA, ARCH, or GARCH models. These models are not pursued in most longitudinal studies due to small samples in T, which render consistent estimation infeasible.

2.2.2. Micro dynamics

Treating the time components (τ_t) as parameters in estimation renders a micro analysis that depicts the stochastic properties of the y_{ti}'s conditional on the τ_t's. This variation characterizes how individuals sort themselves within cross-sectional distributions over time after removing aggregate or economy-wide effects. Such information reveals how individuals systematically deviate from the aggregate trends measured by the quantities τ from one period to the next.

There are two routes through which specification (2.1) captures the micro sources of dynamics: the measured function f_{ti} that relates current and past y_{ti}'s and x_{ti}'s, and the stochastic process generating the unobserved quantities U_{ti}. Expressed in terms of these micro dynamic components, prototype specification (2.5) becomes:

$$\begin{aligned} &f_{ti}(Y_{ti}, Z_{ti}, X_{ti}, \gamma) = y_{ti} - \pi y_{(t-1)i} - \beta_1 x_{ti} - \beta_2 x_{(t-1)i} - \tau_t, \\ &U_{ti} = v_{ti}. \end{aligned} \tag{2.7}$$

Thus, f_{ti} incorporates features of the distributed lag relationships involving measured variables, along with the time effects estimated as parameters. The error U_{ti} captures the time series properties of disturbances. If one were instead to consider specification (2.6), then the error becomes $U_{ti} = \tau_{2t} v^*_{ti}$ which depends directly on time effects and may, as a consequence, exhibit heteroscedasticity properties over time in addition to its autocorrelation attributes.

The remainder of this section discusses a rich array of empirical specifications for modeling both f_{ti} and the intertemporal properties of U_{ti}. The particular variety of model introduced in a longitudinal analysis to characterize individual variation fundamentally depends on the character of the dependent variables. When variables are

discrete, duration or competing risk models are popular candidates. When variables are censored or truncated, researchers commonly specify complete distributional assumptions combining continuous and discrete variables and carry out maximum likelihood procedures. This chapter primarily focuses on estimation methods applicable when the y_{ti}'s are continuously distributed with τ treated as fixed. The analysis covers two distinct types of empirical specifications devised to summarize the micro dynamic properties of y: (i) relationships that link the moments of y determining its autocorrelation structure, and (ii) empirical formulations that describe the evolution of the quantiles of y over time.

2.3. Dynamic simultaneous equation models

Starting with flexible specifications for the function f_{ti} in (2.1) designed to model the intertemporal moments of y, a popular formulation consists of a structural equation from a DSEM (dynamic simultaneous equation model), such as

$$\Pi(L)y_{ti} = \Psi(L)z_{ti} + B(L)x_{ti} + U_{ti}, \tag{2.8}$$

where $\Pi(L) \equiv \sum_{j=0}^{n} \Pi_j L^j$ is a finite-order lag polynomial with $\Pi_0 = 1$, and $\Psi(L) \equiv \sum_{j=0}^{r} \Psi_j L^j$ and $B(L) \equiv \sum_{j=0}^{s} B_j L^j$ are row vectors of finite-order lag polynomials. Written in terms of (2.1), this DSEM implies the specification

$$f_{ti}(Y_{ti}, Z_{ti}, X_{ti}, \gamma) = \Pi(L)y_{ti} - \Psi(L)z_{ti} - B(L)x_{ti}.$$

An alternative representation of (2.8) is

$$y_{ti} = -\sum_{j=1}^{n} \Pi_j y_{(t-j)i} + \sum_{j=0}^{r} \Psi_j z_{(t-j)i} + \sum_{j=0}^{s} B_j x_{(t-j)i} + U_{ti},$$

which may be expressed compactly as

$$y_{ti} = Y'_{(t-1)i}\pi + Z'_{ti}\psi + X'_{ti}\beta + U_{ti}, \tag{2.9}$$

where the vectors $Y_{(t-1)i}$, Z_{ti}, and X_{ti} are defined by (2.2), and the parameter vectors π, ψ and β incorporate coefficients included in Π, Ψ, and B, respectively. With T^* denoting the total number of time periods supplied by the panel data source, $T \equiv T^* - \max(n, r, s)$ is the number of periods for which there are data on all the variables appearing in Equation (2.9). Period 1 in this discussion refers to the first period in which data are available, so $t = 1, \ldots, T$.

Combining observations on Equation (2.9) for a given individual into a single system creates a model that is particularly useful for the analysis of panel data. Stacking these observations in descending order starting with the last period yields

$$\begin{pmatrix} y_{Ti} \\ \vdots \\ y_{1i} \end{pmatrix} = \begin{pmatrix} Y'_{(T-1)i} \\ \vdots \\ Y'_{0i} \end{pmatrix} \pi + \begin{pmatrix} Z'_{Ti} \\ \vdots \\ Z'_{1i} \end{pmatrix} \psi + \begin{pmatrix} X'_{Ti} \\ \vdots \\ X'_{1i} \end{pmatrix} \beta + \begin{pmatrix} U_{Ti} \\ \vdots \\ U_{1i} \end{pmatrix},$$

which, in matrix notation, is

$$y_i = Y_i\pi + Z_i\psi + X_i\beta + U_i, \tag{2.10}$$

$i = 1, \ldots, N$. The disturbances U_{ti} of Equation (2.9) may be autocorrelated over time, but they are assumed to be independently distributed over individuals after the removal of common period effects achieved by including time dummies or polynomials in time among the exogenous variables. Denote the variance–covariance matrix of U_i by $\Theta = E\{U_i U_i'\}$. With distributed lag structures common across individuals, the panel data source offers N independent sets of T time series observations with which to estimate the parameters of Equation (2.9).

Equation (2.8) provides a framework for considering a wide variety of distributed lag relationships among the elements of Y, Z, and X using panel data, including infinite order schemes. The assumption that the lag polynomials $\Pi(L)$, $\Psi(L)$, and $B(L)$ are of finite order is not as restrictive as it may at first appear. One can derive a specification in the form of (2.8) for any infinite-order distributed lag relationship that can be written as a ratio of finite order lag polynomials. Such lag schemes, known as rational distributed lags, admit flexible weighting patterns on past variables and contain many well-known schemes as special cases. Analyzing rational distributed lag schemes using specifications (2.8) or (2.9) can imply nonlinear restrictions relating the coefficients of the polynomials $\Pi(L)$, $\Psi(L)$, and $B(L)$. The estimation procedures developed below permit such restrictions. While imposing these constraints may yield increased efficiency in estimation, one can construct less efficient estimates of infinite-order lag structures using unconstrained estimates and the formulas implied by $\Psi(L)/\Pi(L)$ and $B(L)/\Pi(L)$.[3]

The desirability of imposing "smoothness" restrictions of the sort implied by a rational distributed lag structure has been questioned in the time series literature,[4] and it is natural to question the value of such restrictions in a longitudinal analysis as well. In contrast to a time series analysis, one can completely relax these smoothness restrictions in a panel data setting and test the constraints implied by a particular rational distributed lag scheme before accepting it as a specification. The main implication of assuming that a DSEM characterizes distributed lag relationships among measured variables is the imposition of constraints across equations associated with different time periods for a given individual. Inspection of model (2.10) reveals that these constraints translate into an equality restriction that requires π, ψ and β to be constant across equations. It is

[3] Consider, for example, the construction of estimates for the coefficients of an infinite order rational distributed lag that relates Y_{ti} to a single exogenous variable K_{ti}. According to Equation (2.8), this distributed lag $\delta(L) = \sum_{j=0}^{\infty} \delta_j L^j$ is given by $\delta(L) = B(L)/\Pi(L)$. The result $\Pi(L)\delta(L) = B(L)$ implies formulas for the δ_j's. In particular, equating the coefficients associated with each L^j term in the polynomial $\Pi(L)\delta(L)$ to the corresponding coefficient in the polynomial $B(L)$ yields a set of difference equations that can be solved for δ_j, $j = 0, \ldots,$ given estimates of the Π_j's and the B_j's.

[4] See, for example, Sims (1974) for a discussion of this issue.

straightforward to relax this equality restriction when estimating the system of equations given by (2.10) and to test whether it can be accepted for the longitudinal dataset under consideration.

While specification (2.9) presumes that a researcher wishes to analyze only a single structural equation per period, it is straightforward to modify this specification to permit analysis of a multicolumn vector and of the coefficients Π_j, Ψ_j and B_j as matrices of parameters.

2.4. Modeling dynamics through error structures

In many applications, modeling the autocorrelation properties of disturbances is an important component of a panel data analysis. Indeed, it is the focus of most longitudinal studies in the empirical literature concerned with characterizing the dynamic aspects of an individual's wages or earnings.[5] Besides providing a framework for summarizing the intertemporal properties of variables, the introduction of a stochastic process for disturbances can create a statistical model that may be used for prediction outside the sample period. In the case of simultaneous equations, its inclusion can aid in securing the identification of structural parameters.

2.4.1. Addition of other error components

Many longitudinal analyses combine pure autoregressive or pure moving-average error terms with permanent components and random trend components to model the intertemporal correlation pattern of disturbances. Thus, U_{ti} in Equation (2.8) becomes

$$U_{ti} = \phi_{1i} + \phi_{2i} t + v_{ti}, \tag{2.11}$$

where v_{ti} is now generated by the ARMA process (2.3) with either $a(L) = 1$ or $m(L) = 1$, and ϕ_{1i} and ϕ_{2i} are time-invariant random components distributed independently across individuals.[6]

For those procedures in the subsequent discussion that provide for the direct estimation of either autoregressive or moving-average coefficients contained in the lag polynomials $a(L)$ and $m(L)$, the difference disturbances $(1 - L)v_{ti}$ or $(1 - L)^2 v_{ti}$ should be used in place of the U_{ti}'s when either ϕ_{1i} or ϕ_{2i} are present. In particular, if only the permanent component ϕ_{1i} is admitted (i.e., $\phi_{2i} = 0$), then first-differencing

[5] See, for example, Hause (1977), Lillard and Willis (1978), Baker and Solon (2003), Altonji and Dunn (2000).

[6] The most popular specification is one that combines a permanent component with a pure autoregressive scheme. David (1971), Hause (1977, 1980), Lillard and Willis (1978), and Lillard and Weiss (1979) are examples of studies that estimate first-order schemes (i.e., $p = 1$ and $q = 0$ in (2.3)); Ashenfelter (1978) considers higher-order autoregressive processes. Friedman and Kuznets (1945, p. 353) estimate a first-order moving-average scheme (i.e., $p = 0$ and $q = 1$ in (2.3)) with a permanent component; Hause (1977) considers higher-order moving-average processes. MaCurdy (1982a) considers a mixture of an autoregressive and a moving-average process.

Equation (2.8) creates a new error $(1 - L)U_{ti}$ that follows an ARMA structure of the sort given by (2.3). If ϕ_{2i} is also admitted, then second-differencing changes the specification of the ARMA process for the disturbances in a known way and introduces no new parameters. The same is true if a DSEM specification happens to describe the relationships involving the measured variables. Thus, it is possible to construct a full set of estimates for the coefficients of the polynomials $\Pi(L)$, $\Psi(L)$, $B(L)$, $a(L)$, and $m(L)$ using only the parameter estimates of the differenced equations.

Alternatively, if the error components ϕ_{1i} and ϕ_{2i} are assumed to be independent of the ν_{ti}'s, as is often maintained in longitudinal studies, it is straightforward to modify the specification of $\Theta = E\{U_i U_i'\}$ developed below to reflect the influence of ϕ_{1i}, ϕ_{2i}, or both. This adjusted specification of Θ can then be used in the estimation procedures proposed in later sections.

Yet another set of error structures implemented in a longitudinal setting describes disturbances as taking the form

$$U_{ti} = \lambda_t' \phi_i + \nu_{ti}, \quad t = 1, \ldots, T, \ i = 1, \ldots, N, \tag{2.12}$$

where $\lambda_t' = (\lambda_{t1}, \ldots, \lambda_{tk})$ is a vector of factor-loading coefficients (which may or may not be known); and $\phi_i' = (\phi_{1i}, \ldots, \phi_{ki})$ is a vector of individual-specific time-invariant disturbances with

$$E(\phi_i \phi_j') = \begin{cases} \Omega & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

According to (2.12), U_{ti} equals a weighted sum of two error structures: a k-dimensional "factor" model consisting of a sum of correlated individual-specific errors ϕ_j; and an individual-time-specific error ν_{ti} which is distributed independently of ϕ_i and may be serially independent over time or follow an ARMA(p, q) process. This error structure admits a wide variety of autocorrelation patterns with a minimal number of parameters.

As in the case of (2.11), it is elementary to modify the specification of $\Theta = E\{U_i U_i'\}$ developed below to account for the influence of the factor components ϕ_i appearing in (2.12). This adjusted specification of Θ can then be used in the estimation procedures proposed in later sections. Alternatively, one can often transform (2.12) into a form that eliminates the ϕ components, and one can estimate the transformed error structure utilizing approaches similar to the differenced specifications outlined above. For example, in the most common formulations, the factor loading λ_t and error component ϕ_i are each scalars. In such cases, one can divide (2.12) by λ_t and apply first-differencing to eliminate ϕ_i. The resulting specification introduces an individual-time-specific error, ν_{ti}/λ_t, which may be heteroscedastic over time. In terms of specification (2.4), this implies the white noise errors ε_{ti} will be nonstationary.

2.4.2. *Admitting nonstationarity in longitudinal analysis*

Permitting the errors ε_{ti} to be heteroscedastic over time gives rise to no conceptual difficulties in analyses of panel data. The variance–covariance parameters Σ_{ti} in (2.4)

differ across t but remain constant over individuals i. In standard time series analysis, this sort of nonstationarity does not necessarily create estimation problems, but it does require an explicit parameterization of the suspected form of the heteroscedasticity that avoids an incidental parameters problem. In the case of panel data, however, estimation procedures allow for arbitrary forms of serial heteroscedasticity.

A second form of nonstationarity permitted in panel data analysis provides for the coefficients of the lag polynomials $a(L)$ and $m(L)$ to vary arbitrarily across periods so that there are new sets of autoregressive and moving-average parameters for each t. Specification (2.3) incorporated such options by including t subscripts on the lag polynomials $a_t(L)$ and $m_t(L)$; testing of common coefficients in a longitudinal analysis setting is discussed below.

A third source of nonstationarity readily admitted in panel data analysis involves relaxing the requirement for the roots of the autoregressive lag operator $a(L)$ to lie outside the unit circle. Thus, it is possible to consider such error processes as random walks when using longitudinal data. In a time series analysis the existence of such nonstationarity has significant consequences on the asymptotic properties of estimators, but in the case of panel data, where asymptotic results rely on a large number of individuals rather than a large number of time periods, its inclusion has no such effects.

A fourth form of nonstationarity found in panel data studies comes about due to the influence of initial conditions associated with the starting values of an ARMA model, a set of conditions that differ across individuals. Section 4 considers this topic in detail.

To highlight key ideas, Sections 4 and 7 focus on the modeling and estimation problems encountered when assuming that a common ARMA process generates the disturbances U_{ti} over time and across individuals, with the white noise errors ε_{ti} assumed to be stationary and the same for all individuals. Such error structures admit a wide variety of time series aspects of panel data with a minimal number of parameters. Moreover, in longitudinal analyses this class of error specifications has performed well in describing the intertemporal features of the data. The subsequent discussion also covers modifications needed to accommodate each of the sources of nonstationarity outlined above.[7]

2.5. *Dynamic quantile regressions*

Rather than introducing specifications describing the evolution of the mean and the autocorrelation structure of the y_{ti}'s, suppose one instead wishes to characterize the micro dynamic properties of these dependent variables by modeling the intertemporal features of the conditional quantiles of y. To interpret prototype equation (2.5) in this context, assume the error v_{ti} in this specification is distributed independently both over time and across individuals; so, $v_{ti} = \varepsilon_{ti}$, the white noise error term specified in (2.4). The autoregressive coefficient π characterizes the dynamic properties of wages after removing

[7] Baltagi (2002) surveys recent developments of panel data methods for estimating parameters in the presence of several varieties of nonstationarity popular in the times series literature.

trends, and one can readily generalize this prototype specification to incorporate an autoregressive structure with multiple lags, as analogous to the DSEM specified by (2.8).

A conventional autoregressive formulation of (2.5) invokes the moment restriction

$$E(\nu_{ti}|Y_{(t-1)i}, X_{ti}) = 0, \tag{2.13}$$

where $Y_{(t-1)i}$ signifies the past wages $y_{(t-1)i}, \ldots, y_{(t-k)i}$. This condition implies that (2.5) characterizes how the first moment of the Markov distribution of y_{ti} conditional on $Y_{(t-1)i}$ and X_{ti} evolves over time. One applies least squares or generalized least squares methods to estimate the parameters of such formulations, suitably adjusting for heteroscedasticity or correlation in an individual's errors when appropriate.

Alternatively, one can associate relation (2.5) with an autoregressive formulation of the κth percent quantile (or percentile) of the Markov distribution of y_{ti} by imposing the restriction

$$q_\kappa(\nu_{ti}|Y_{(t-1)i}, X_{ti}) = 0, \tag{2.14}$$

where $q_\kappa(\,)$ designates the κth percent quantile of the distribution of ν_{ti} conditional on $Y_{(t-1)i}$ and X_{ti}, where $\kappa \in (0, 100)$. When $\kappa = 50$, Equation (2.5) determines how the conditional median of y_{ti} evolves over time. Conceptually, the application of LAD procedures would produce consistent estimates of the autoregressive coefficients appearing in (2.5). Such relations have only rarely been estimated in a panel data context, but specifying variants of (2.14) for several values of κ offers a parsimonious and flexible framework for describing micro dynamic relationships.

Section 5 summarizes a class of quantile estimators for systems of simultaneous equation models comprised of time series observations on $f_{ti}(Y_{ti}, Z_{ti}, X_{ti}, \gamma)$ that provides a flexible and noncumbersome procedure for estimating parameters of dynamic relationships of the sort specified by (2.5). In essence, assuming specifications for the quantiles of structural error distributions conditional on exogenous or predetermined instruments, the estimators formulate these conditional quantiles into moment conditions capable of being estimated within a conventional nonlinear instrumental variable or MM (method of moments) framework. This apparatus matches the sample analog of the conditional quantiles against their population value, employing a smoothing procedure familiar in various problems found in nonparametric inference and simulation estimation. The analysis applies standard arguments to demonstrate consistency and asymptotic normality of the resulting Smoothed MM Quantile estimator.

3. Basic estimation concepts and challenges

Much is known about the estimation of nonlinear simultaneous equations of the sort encountered in longitudinal data analyses, and this section outlines the key results. The discussion opens with a brief summary of the "method of moments" (MM) estimation framework, which also goes by the names of "generalized method of moments" (GMM),

"nonlinear instrumental variables" (NIV), and "minimum chi-squared" estimation.[8] In addition to reviewing the central asymptotic results exploited in the MM estimation approach, the discussion also highlights the procedures for selecting optimal instrumental variables and for testing joint hypotheses.

While this apparatus offers a comprehensive framework for estimating a wide variety of models, one encounters several challenges in applying these methods in a panel data context. One challenge, which has been addressed fully in the literature, concerns how to exploit predetermined variables – variables that are endogenous in some equations but not in others – as instrumental variables in estimation. More demanding undertakings involve computational complexities that arise when one incorporates large numbers of equations in estimation with intricate nonlinearities, as well as development of empirical specifications enabling estimation of the dynamic structure describing error processes. Still more formidable challenges concern how to use weights to account for stratified samples that are a part of all longitudinal surveys, and how to carry out estimation with unbalanced datasets – samples with an uneven number of time periods and possibly nonoverlapping time periods of data available for individual sample members. The second part of this section provides an overview of these challenges, whereas Sections 4 through 6 lay out specific approaches for dealing with each challenge in a longitudinal data setting.

3.1. *Overview of method of moments estimation procedures*

Suppose one is interested in calculating a consistent estimate of the "true" value of a $p \times 1$ parameter vector γ that is an unknown determinant of the distribution generating a random vector Y. Denote this true value as γ_0, and let Y_i and X_i, $i = 1, \ldots, N$, denote N observations on Y and on a vector of measured characteristics X. The Y_i's are assumed to be independently distributed across observations after conditioning on the X_i's, or when these characteristics are treated as known constants.

3.1.1. *Method-of-moments estimators*

The MM approach offers a general procedure for estimating the parameters γ_0 in large samples. To characterize this class of estimators, let $\ell_i(\gamma) \equiv \ell(\gamma, M_i)$, $i = 1, \ldots, N$, represent a $b \times 1$ vector of known functions with $b \geqslant p$ where the vector M_i includes elements of Y_i and X_i. Consider the system of equations

$$L_N(\gamma) \equiv \frac{1}{N} \sum_{i=1}^{N} \ell_i(\gamma) = 0. \tag{3.1}$$

[8] Many have contributed to the development of this estimation methodology. Most notably, Sargan (1958, 1961) initiated the study of this class of estimators, and Amemiya (1974, 1975, 1977) and Hansen (1982) substantially generalized and expanded these methods to create the broad framework summarized in this section. Handbook chapters by Amemiya (1983) and Manski (1994) present valuable overviews and alternative presentations of these approaches.

Assuming each ℓ_i possesses a sufficiently well-behaved distribution and the ℓ_i's are chosen so that $\lim_{N\to\infty} E(L_N(\gamma_0)) = 0$, one can show that setting $\gamma = \gamma_0$ solves (3.1) in the sense that $L_N(\gamma_0)$ converges in probability to zero as the sample size goes to infinity. Identification of γ requires the existence of a unique solution to (3.1), which requires the Jacobian of L_N to have the appropriate rank, as dictated by the implicit function theorem.[9] The MM estimation approach maintains that the matrices of first partials $\frac{\partial \ell_i}{\partial \gamma'}$, $i = 1, \ldots, N$, exist with each element uniformly continuous in γ. Denote the average of these partials by $S_N(\gamma) \equiv \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell_i}{\partial \gamma'}$, assumed to possess full column rank, and define the matrix $V_N(\gamma) \equiv \frac{1}{N}\sum_{i=1}^{N} \ell_i(\gamma)\ell_i(\gamma)'$ as an average of outer products. By maintaining further assumptions guaranteeing satisfaction of a set of regularity conditions, one can demonstrate that computing a solution to (3.1) yields a strongly consistent estimate for γ_0 that is asymptotically normally distributed.[10] To derive the asymptotic results cited below, the distributions associated with the ℓ_i's and the matrices of first partials cannot have too much weight in the tails.[11]

When the number of equations in (3.1) used to compute estimates surpasses the number of parameters, there seldom exists a value for γ that solves all equations making up (3.1) exactly in finite samples. Thus, one requires a weighting scheme for comparing the errors obtained in solving the various equations. A standard approach is to compute a value $\tilde{\gamma}$ to minimize the quadratic form

$$L_N(\gamma)' H_N L_N(\gamma), \tag{3.2}$$

where H_N is a positive definite symmetric matrix for all N (including its probability limit as $N \to \infty$). When the number of equations exceeds the number of parameters, the form of H_N determines the relative tradeoffs in solving (3.1), which in turn defines the estimator $\tilde{\gamma}$.[12] Essentially any $\tilde{\gamma}$ that minimizes (3.2) yields a strongly consistent estimator for γ_0 that is asymptotically normally distributed as follows:

$$\sqrt{N}(\tilde{\gamma} - \gamma_0) \xrightarrow{d} N\Big(0, \operatorname*{plim}_{N\to\infty}\Big[\big[\widetilde{S}_N' H_N \widetilde{S}_N\big]^{-1}\big[\widetilde{S}_N' H_N \widetilde{V}_N H_N \widetilde{S}_N\big]\big[\widetilde{S}_N' H_N \widetilde{S}_N\big]^{-1}\Big]\Big),$$

[9] In many applications, one cannot rule out the possibility that values of γ other than γ_0 may also satisfy (3.1) in the limit. One can, however, easily resolve this issue for the estimation problems considered below, and for simplicity this analysis assumes the solution to (3.1) is unique.

[10] To prove consistency and asymptotic normality of the solution to $L_N(\gamma) = 0$, the convergence of L_N, S_N, and V_N to their respective limits must be uniform in γ, and this assumption is maintained throughout the discussion. Chapter 4 in Amemiya (1985) provides detailed definitions of several forms of uniform convergence.

[11] Letting ℓ_{ji} and s_{jki} denote the j and (j,k) elements of ℓ_i and $\frac{\partial \ell_i}{\partial \gamma'}$, respectively, sufficient conditions restricting the tails of distributions are: $E|\ell_{ji}|^{2+\delta_2} \leqslant C_1 < \infty$ and $E|s_{jki}|^{1+\delta_1} \leqslant C_2 < \infty$ for some $\delta_1, \delta_2 > 0$ and all $\gamma \in \Gamma$.

[12] Ordinary least squares derives an estimate for γ by minimizing the sum of squared errors associated with the b equations appearing in (3.1), which implies setting $H_N = I$ ($\equiv$ identity matrix), corresponding to minimizing the quantity $L_N(\gamma)' L_N(\gamma)$.

where $\widetilde{S}_N \equiv S_N(\tilde{\gamma})$ and $\widetilde{V}_N \equiv V_N(\tilde{\gamma})$. Thus, the approximate distribution for $\tilde{\gamma}$ in large samples is

$$\tilde{\gamma} \dot{\sim} N\left(\gamma_0, \frac{1}{N}\left[\left[\widetilde{S}_N' H_N \widetilde{S}_N\right]^{-1}\left[\widetilde{S}_N' H_N \widetilde{V}_N H_N \widetilde{S}_N\right]\left[\widetilde{S}_N' H_N \widetilde{S}_N\right]^{-1}\right]\right). \tag{3.3}$$

The efficiency of the estimator $\tilde{\gamma}$ depends on the choice of H_N.

3.1.2. Generalized least squares

According to generalized least squares theory, selecting H_N in (3.2) to be a matrix that is proportional to the inverse of the variance–covariance matrix of $L_N(\gamma_0)$ leads to the most efficient parameter estimate obtained by minimizing (3.2). Such a choice is $H_N = [E(V_N(\gamma_0))]^{-1}$, where this expression relies on the relation

$$E\left[L_N(\gamma_0)L_N(\gamma_0)'\right] \xrightarrow{p} \frac{1}{N}E\left[V_N(\gamma_0)\right],$$

following from the independence of observations ℓ_i with the notation $\xrightarrow{p}$ designating convergence in probability. The matrix $E(V_N(\gamma_0))$ is unknown, but as with many generalized least squares analyses, a consistent estimate for this matrix is easily constructed and the asymptotic properties of estimators are unaffected if one substitutes this consistent estimate for the true value of the matrix. Accordingly, when computing an estimate for γ_0, one sacrifices no estimation efficiency by instead minimizing the quadratic-form distance function

$$C(\gamma) \equiv L_N(\gamma)' \widetilde{V}_N^{-1} L_N(\gamma), \tag{3.4}$$

where $\widetilde{V}_N \equiv V_N(\tilde{\gamma})$ with $\tilde{\gamma}$ representing any consistent estimate for γ_0, implying $\widetilde{V}_N \xrightarrow{p} \lim_{N\to\infty} E(V_N(\gamma_0))$. Let $\hat{\gamma}$ denote that value of γ minimizing (3.4).

The asymptotic properties of the estimator $\hat{\gamma}$ follow from (3.3). With $H_N = \widetilde{V}_N^{-1}$, we have $\hat{\gamma} \xrightarrow{s} \gamma_0$ and

$$\sqrt{N}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N\left(0, \left[S(\gamma_0)' V^{-1}(\gamma_0) S(\gamma_0)\right]^{-1}\right),$$

where the matrices $S(\gamma_0)$ and $V(\gamma_0)$ respectively denote the probability limits of $S_N(\gamma_0)$ and $V_N(\gamma_0)$. Thus, the approximate distribution for $\hat{\gamma}$ in large samples is

$$\hat{\gamma} \dot{\sim} N\left(\gamma_0, \frac{1}{N}\left[\widehat{S}_N' \widetilde{V}_N^{-1} \widehat{S}_N\right]^{-1}\right), \tag{3.5}$$

where $\widehat{S}_N \equiv S_N(\hat{\gamma})$.

Distribution formula (3.5) also applies when equation system (3.1) yields a just-identified solution for the parameters γ. In such instances, the number of equations in (3.1) equals the number of elements in γ. Since the system of equations $L_N = 0$ alone fully defines γ, the choice of H_N in (3.2) is irrelevant in the calculation of $\tilde{\gamma}$. In

such circumstances, the asymptotic distribution simplifies to

$$\tilde{\gamma} \dot{\sim} N\left(\gamma_0, \frac{1}{N}\widetilde{S}_N^{-1}\widetilde{V}_N\widetilde{S}_N'^{-1}\right). \tag{3.6}$$

One finds this specification of the large-sample distribution reported for many estimators.[13]

3.1.3. Instrumental variable estimators

Longitudinal empirical models of the sort outlined in Section 2 comprise special cases of the following system of nonlinear simultaneous equations:

$$f_i(\gamma_0) = \xi_i, \quad \text{where } f_i \equiv f_i(\gamma) \equiv \begin{pmatrix} f_{Ti} \\ \vdots \\ f_{1i} \end{pmatrix} \equiv \begin{pmatrix} f_T(M_i, \gamma) \\ \vdots \\ f_1(M_i, \gamma) \end{pmatrix} \text{ and } \xi_i \equiv \begin{pmatrix} \xi_{Ti} \\ \vdots \\ \xi_{1i} \end{pmatrix} \tag{3.7}$$

where the $f_{ti}, i = 1, \ldots, N, t = 1, \ldots, T$, are vectors of known functions, the column vector M_i represents the ith observation on measured variables, γ_0 denotes the true value of the parameter vector γ that generates the sample under consideration, and ξ_i represents an error vector which is distributed independently across observations with $E\{\xi_i\} = 0$. Designate Q_i as a vector whose elements are functions of instrumental variables, and which is presumed in the subsequent discussion to always include a constant. In conventional simultaneous nonlinear equations, one maintains the assumption that $E\{\xi_i | Q_i\} = 0$, so we have conditional-first-moment independence of ξ_i and Q_i.

A formulation for ℓ_i's in the MM framework satisfying the properties needed for consistent estimation of γ_0 takes the form

$$\ell_i = \begin{pmatrix} (f_{Ti} \otimes Q_{Ti}) \\ \vdots \\ (f_{1i} \otimes Q_{1i}) \end{pmatrix} \equiv \Delta_i f_i, \tag{3.8}$$

where the operator $\otimes$ designates a matrix Kronecker product,[14] and the matrix Δ_i is given by

$$\Delta_i = \begin{pmatrix} (I_{n_T} \otimes Q_{Ti}) & 0 & \cdots & 0 & 0 \\ 0 & (I_{n_{T-1}} \otimes Q_{(T-1)i}) & \cdots & \cdot & \cdot \\ \cdot & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & (I_{n_2} \otimes Q_{2i}) & 0 \\ 0 & 0 & \cdots & 0 & (I_{n_1} \otimes Q_{1i}) \end{pmatrix}, \tag{3.9}$$

[13] Prominent examples include least squares procedures, wherein (3.1) corresponds to the "moment conditions".

[14] The Kronecker is defined as $(f_{ji} \otimes Q_{ki}) = \begin{pmatrix} f_{1ji}Q_{ki} \\ \vdots \\ f_{mji}Q_{ki} \end{pmatrix}$, where $f_{ji} = \begin{pmatrix} f_{1ji} \\ \vdots \\ f_{mji} \end{pmatrix}$.

with I_{n_j} denoting an identity matrix of dimension n_j. If there exist the same number of structural equations in each period, then $n_T = \cdots = n_1$. This formulation of ℓ_i allows different functions of the instrumental variables to be applied to each set of equations f_{ti} in estimating the parameter vector γ. The distance function (3.4) multiplied by N^2 takes the form

$$N^2 C(\gamma) \equiv \left[\Sigma_i \ell_i(\gamma)'\right]\left[\frac{1}{N}\Sigma_i\left(\tilde{\ell}_i\tilde{\ell}_i'\right)\right]^{-1}\left[\Sigma_i \ell_i(\gamma)\right],$$

where $\tilde{\ell}_i$ is a consistent estimate of ℓ_i. This may also be written as:

$$\left[\Sigma_i f_i' \Delta_i'\right]\left[\frac{1}{N}\Sigma_i\left(\Delta_i\tilde{\xi}_i\tilde{\xi}_i'\Delta_i'\right)\right]^{-1}[\Sigma_i \Delta_i f_i], \tag{3.10}$$

where $\tilde{\xi}_i$ is a consistent estimate of ξ_i. According to (3.5), this instrumental variable estimator possesses the large sample multivariate normal distribution:

$$\hat{\gamma} \mathrel{\dot{\sim}} N\left(\gamma_0, \left[\left[\Sigma_i \frac{\partial f_i'}{\partial \gamma}\bigg|_{\hat{\gamma}} \Delta_i'\right]\left[\Sigma_i\left(\Delta_i\tilde{\xi}_i\tilde{\xi}_i'\Delta_i'\right)\right]^{-1}\left[\Sigma_i \Delta_i \frac{\partial f_i}{\partial \gamma'}\bigg|_{\hat{\gamma}}\right]\right]^{-1}\right), \tag{3.11}$$

where $\frac{\partial f_i}{\partial \gamma'}|_{\hat{\gamma}}$ is a matrix of gradients evaluated at $\hat{\gamma}$.

3.1.4. Optimal choice of instrumental variables

The efficiency of the MM estimator depends on the selection of the instrumental variables Q_{ti} implemented to estimate the various equations. Assuming the errors ξ_i in system (3.7) have a common variance–covariance structure across individuals, the optimal choice of the functional forms of the instrumental variables – the form producing the most efficient estimates – is given by

$$Q_{ti} = E\left(\frac{\partial f_{ti}}{\partial \gamma_t}\bigg|_{\gamma_0} | X_i\right), \quad t = 1, \ldots, T, \tag{3.12}$$

where the parameter vector γ_t includes the components of γ appearing in f_{ti}. The expectations in (3.12) condition on all the exogenous variables available in the simultaneous equations system (3.7), designated here as X_i.[15] With this formulation for the Q_{ti}'s, minimizing the function (3.4) computes the MM estimate, and its asymptotic distribution is given by (3.11). Considering the class of all MM estimators with Q_{ti} being any function of the instrumental variables, the estimator computed with the Q_{ti} selected according to (3.12) yields the most efficient estimator in this class when errors are homoscedastic.

[15] See Amemiya (1975) for the original demonstration of this result. Formula (3.12) assumes that f_{ti} is a single structural equation. If f_{ti} is a vector, then one forms Q_{ti} by stacking the columns of the matrix $E(\frac{\partial f_{ti}'}{\partial \gamma_t}|_{\gamma_0}|X_i)$.

In general, this choice of Q_{ti} is not directly observed because γ_0 is unknown. However, there is no loss of efficiency by replacing (3.12) with

$$Q_{ti} = E\left(\left.\frac{\partial f_{ti}}{\partial \gamma_t}\right|_{\tilde{\gamma}} | X_i\right), \quad t = 1, \ldots, T, \tag{3.13}$$

where $\tilde{\gamma}$ is a consistent estimate for γ_0, treated as fixed in the calculations of the expectation. A popular procedure for approximating the formulation of Q_{ti} given by (3.13) is to

$$\text{project } \left.\frac{\partial f_{ti}}{\partial \gamma_t}\right|_{\tilde{\gamma}} \text{ on functions of elements of } X_i. \tag{3.14}$$

This is usually accomplished by regressing components of $\frac{\partial f_{ti}}{\partial \gamma_t}|_{\tilde{\gamma}}$ on functions of X_i. One then uses the fitted values from this regression to serve as Q_{ti}.

3.1.5. Testing procedures

Two approaches are popular for testing hypotheses involving general forms of nonlinear restrictions relating the elements of γ. Consider the null and alternative hypotheses

$$H_0\colon\ r(\gamma) = 0 \quad \text{versus} \quad H_a\colon\ r(\gamma) \neq 0, \tag{3.15}$$

where $r(\gamma)$ is a $q \times 1$ vector of known functions specifying the q restrictions linking the components of γ. One form of test relies on a Wald statistic, and a second exploits a likelihood-ratio type statistic.

To construct a Wald statistic, define the partial derivative matrix and its corresponding estimate as

$$R(\gamma) \equiv \frac{\partial r}{\partial \gamma'} \quad \text{and} \quad \widehat{R} = R(\hat{\gamma}).$$

Assuming that H_0 contains no redundant restrictions, one can formulate a matrix R that possesses full row rank. If H_0 is true, then

$$N r(\hat{\gamma})' \left[\widehat{R}\left[\widehat{S}_N' \widetilde{V}_N^{-1} \widehat{S}_N\right]^{-1} \widehat{R}'\right]^{-1} r(\hat{\gamma}) \mathrel{\dot{\sim}} \chi_q^2; \tag{3.16}$$

that is, under the null hypothesis, the Wald statistic is approximately distributed, for sufficiently large N, according to a chi-squared distribution with q degrees of freedom.

A comparison of the optimized values of the distance functions (3.4) when assuming the null and alternative hypotheses provides another basis for testing H_0. If H_0 is true, then it can be shown that

$$\left[N \min_{\gamma\colon\, r(\gamma)=0} C(\gamma)\right] - \left[N \min_{\{\gamma\}} C(\gamma)\right] \mathrel{\dot{\sim}} \chi_q^2, \tag{3.17}$$

where $\widetilde{V}_N$ in (3.4) is kept constant when minimizing $C(\gamma)$ under H_0 and H_a. The first term in (3.17) computes $C(\gamma)$ imposing the constraints $r(\gamma) = 0$, whereas the second term minimizes $C(\gamma)$ invoking no restrictions. Thus, under the null hypothesis, the

difference in optimized distance functions used to calculate estimates is approximately distributed, for sufficiently large N, according to a chi-squared distribution with q degrees of freedom.

3.2. Challenges

One encounters a variety of problems in implementing the above MM framework when estimating panel data specifications. Some are easily overcome, such as incorporating predetermined variables as instruments in estimation. Others can become particularly troublesome, such as avoiding computational difficulties with large systems, formulating empirical specifications to estimate the correlation pattern of sophisticated error structures, and using weights and unbalanced samples in estimation. The following discussion briefly reviews these challenges, while the next sections present options for overcoming the various problems.

3.2.1. Simultaneous equations with predetermined variables

When estimating time series models within a simultaneous equation system, it is often necessary or desirable to exploit the fact that certain variables can be considered predetermined for a subset of the equations. This involves using these variables as instruments in estimating some equations, while treating these same variables as endogenous in others. Previous literature has fully addressed methods for exploiting predetermined variables in the estimation of longitudinal models.[16]

To convey the essential ideas underlying these methods, consider a system of structural equations

$$g_i \equiv \begin{pmatrix} g_{Ti} \\ \vdots \\ g_{1i} \end{pmatrix} \equiv \begin{pmatrix} g_{Ti}(M_{Ti}, \gamma_0) \\ \vdots \\ g_{1i}(M_{1i}, \gamma_0) \end{pmatrix} = \begin{pmatrix} \nu_{Ti} \\ \vdots \\ \nu_{1i} \end{pmatrix} \equiv \nu_i, \tag{3.18}$$

where $g_{ti}(\cdot)$ and ν_{ti} are directly analogous to $f_{ti}(\cdot)$ and ξ_{ti} appearing in model (3.7). Interpret g_{ti} as a structural equation associated with time period t. Suppose that the variables included in the column vector P_{ti} (a subset of M_{ti}) are predetermined for this equation. While a "predetermined property" often refers only to covariance restrictions (e.g., $E\{P_{ti}\nu_{ti}\} = 0$), much of the discussion below interprets predetermined as implying that all the elements of the vector P_{ti} are distributed independently of the error ν_{ti}, but not necessarily of the errors ν_{ki}, $k < t$. The exogenous variables of model (3.18) –

[16] The Handbook chapter by Arellano and Honoré (2001) presents an extensive review of these methods. Studies exploiting predetermined variables as instrumental variables in the estimation of longitudinal models include Anderson and Hsiao (1982), Bhargava and Sargan (1983), Amemiya and MaCurdy (1986), Holtz-Eakin, Newey and Rosen (1988), Arellano and Bond (1991), Keane and Runkle (1992), and Arellano and Bover (1995).

grouped into the vector X_i – are assumed to be distributed independently of the errors $v_{ti}, t = 1, \ldots, T$.

Efficient estimation of this system of equations requires the use of the P_{ti}'s as instrumental variables for the appropriate equations. More importantly, for some models, predetermined variables are the only source of instruments, which makes it necessary to devise an estimation procedure that exploits these variables.

An obvious method for using the MM procedure described above to calculate an estimate for γ_0 utilizing all the available instrumental variables is to create a new structural model in the form of (3.8) by setting

$$\ell_i = \begin{pmatrix} g_{Ti} \otimes \binom{Q_{Ti}}{P_{Ti}} \\ \vdots \\ g_{1i} \otimes \binom{Q_{1i}}{P_{1i}} \end{pmatrix} \equiv \begin{pmatrix} g_{Ti} \otimes Q^*_{Ti} \\ \vdots \\ g_{1i} \otimes Q^*_{1i} \end{pmatrix} \equiv \Delta^*_i g_i, \tag{3.19}$$

with Δ^*_i defined analogously to (3.9). This formulation for ℓ_i satisfies the properties required in the above discussion for minimization of (3.4) to result in a consistent estimate for γ_0 that is asymptotically normally distributed according to distribution (3.5). Accordingly, the distance function (3.10) and the asymptotic distribution (3.11) apply when estimating (3.18) with g_{ti}, $\tilde{v}_i$, Q^*_{ti}, and Δ^*_i replacing f_{ti}, $\tilde{\xi}_i$, Q_{ti}, and Δ_i in (3.10) and (3.11), respectively.

3.2.2. *Optimal instrumental variables with predetermined variables in the MM framework*

As in the case with only exogenous variables, with predetermined variables present the efficiency of the MM estimator depends on the selection of the instrumental variables. Assuming homoscedasticity of errors v_i, a linear transformation of equation system (3.18) puts it into a simpler form for characterizing this efficient estimator. Premultiplying Equation (3.18) by the matrix B that is constructed to be lower triangular with $BE(vv')B' = I$ transforms the model into the form

$$g^\#_i \equiv Bg_i = Bv_i \equiv v^\#_i. \tag{3.20}$$

This linear transformation creates a model with errors possessing the covariance structure $E(v^\# v^{\#\prime}) = I$, while maintaining the particular "predetermined properties" assumed for the P_{ti}'s, with the $v^\#_{ti}$'s serving in place of the v_{ti}'s; namely, the variables P_{ti} are distributed independently of the errors $v^\#_{ti}, \ldots, v^\#_{Ti}$, but not of the errors $v^\#_{1i}, \ldots, v^\#_{(t-1)i}$.

Implementing MM to estimate the coefficients γ computes an estimator to minimize (3.10) with

$$f_i = g^\#_i;$$
$$\Delta_i = \Delta^\#_i \quad \text{with } Q_{ti} = Q^\#_{ti} \text{ for } t = 1, \ldots, T; \quad \text{and}$$
$$I_T = E\{v^\# v^{\#\prime}\} \text{ replaces } \tilde{\xi}_i \tilde{\xi}'_i.$$

With these substitutions, distribution (3.11) specifies the asymptotic distribution of this NIV estimator.

In place of expression (3.12), the optimal choice for $Q_{ti}^{\#}$ accounting for predetermined variables takes the form[17]

$$Q_{ti}^{\#} = E\left(\left.\frac{\partial g_{ti}^{\#}}{\partial \gamma_t}\right|_{\gamma_0} | X_i, P_{ti}\right), \quad t = 1, \ldots, T. \tag{3.21}$$

Considering the class of all NIV estimators with $Q_{ti}^{\#}$ being any functions of the exogenous and predetermined variables available for errors $\nu_{ti}, \ldots, \nu_{Ti}$, the NIV estimator computed with $Q_{ti}^{\#}$ selected according to (3.21) yields the most efficient estimator in this class when errors ν_i are homoscedastic across individuals. Because $Q_{ti}^{\#}$ is not directly observed, a procedure for constructing (3.21) is to

$$\text{project } \left.\frac{\partial g_{ti}^{\#}}{\partial \gamma_t}\right|_{\tilde{\gamma}} \text{ on functions of elements of } X_i \text{ and } P_{ti}, \tag{3.22}$$

where $\tilde{\gamma}$ is any consistent estimate of γ_0. Then use the fitted values of this projection as the optimal $Q_{ti}^{\#}$.

3.2.3. *Specifications providing for estimation of ARMA coefficients and dynamic quantiles*

Often the aim of a longitudinal analysis involves discovering the characteristics of the error structure of a model, and the challenge becomes developing empirical specifications that allow for the estimation of parameters governing either the autocorrelation pattern or the dynamic evolution of the quantiles associated with errors appearing in structural equations. Obviously, if the f_{ti}'s in the above framework only refer to the original structural equations linking measured variables, then the information signaling the dynamic properties of error terms merely shows up as a determinant of the standard errors of coefficients that the above analysis computes in an unrestricted fashion. For the MM framework to be of use in informing researchers about the intertemporal properties of error processes, additional specifications must be formulated for some of the f_{ti}'s that capture the restrictions implied by the proposed error structure. Moreover, these additional f_{ti}'s must be combined with the original structural equations so as to identify all parameters and meet the conditions maintained by the MM framework.

When the disturbances U_{ti} in (2.7) follow a pure AR process, simple and well-known linear transformations of the original equations create the specifications needed

[17] See the appendix of Amemiya and MaCurdy (1986) for the derivation of this optimal choice. The Handbook chapter by Arellano and Honoré (2001) surveys recent developments and generalizations of this specification of optimal instrumental variables, and also covers a variety of interesting refinements in estimation methods.

to estimate the AR coefficients. It is also straightforward to derive the additional specifications needed to estimate parameters associated with a pure MA error process. The task becomes substantially more difficult when a mixed ARMA model describes the intertemporal properties of the U_{ti}'s. In such instances, one must accommodate the consequences of "initial conditions" in specifications, which can be a formidable task, as demonstrated in the next section.

An attractive alternative to estimating moment relationships characterizing the micro intertemporal properties of variables involves using conditional quantile regressions to describe these dynamics. To date, the MM framework outlined above has not been directly applied to estimate parameters of such specifications in a panel data setting. As demonstrated in Section 5, this framework offers a flexible empirical approach for estimating autoregressive specifications of quantile equations.

3.2.4. Potential computational issues

A variety of software routines exist for implementing the above formulation of the MM estimation framework, albeit in some conventional statistical packages one must undertake programming beyond the built-in procedures. In the use of any of these routines, one can encounter considerable computational problems in applying this approach in a panel data setting. Two factors contribute to these difficulties.

First, estimation of longitudinal specifications often results in structures of the ℓ_i's that have large dimensions, leading to potential problems in inverting the matrix $\widetilde{V}_N$ as required to calculate estimates using (3.4) and to compute asymptotic distributions using (3.5). Consider, for example, estimation of the simple linear prototype equation (2.5) using a longitudinal dataset. In conducting this estimation, suppose: (i) the errors ν_{ti} in (2.5) follow a MA(2) process; (ii) the vectors x_{ti} each include 5 exogenous variables that are linearly independent of the other x_{ki}'s, $k \neq t$; (iii) the coefficients π, β_1, and β_2 differ over time; and (iv) a researcher has 10 periods of data along with information on the initial conditions y_{0i} and y_{-1i}. To estimate coefficients of the period-t variant (2.5), available instrumental variables include the exogenous variables x_{ti}, $x_{(t-1)i}$, the predetermined variables $y_{(t-2)i}, \ldots, y_{-1i}$, and time dummies. (The inclusion of the 10 time dummies, of course, identifies the time effects τ.) Thus, joint estimation of all 10 period equations implies a construction of ℓ_i in (3.19) that incorporates at least 165 elements. Moreover, for each period-t equation, all x_{ki} for $k \neq t, t-1$ also constitute valid instrumental variables providing for increased prediction of variation in $y_{(t-1)i}$ beyond that captured in the above list, instruments which if exploited would enhance the efficiency of estimation. This implies existence of an additional 45 ($= 5 \cdot 9$) instrumental variables per equation. Incorporating all these over-identifying variables in estimation would expand ℓ_i by 450 elements. If, instead, a researcher merely employs 4 over-identifying restrictions per equation, then ℓ_i contains more than 200 elements. Consequently, the dimension of the $\widetilde{V}_N$ matrix is over 200×200 in this simple case. While inverting such a matrix is conceptually manageable using familiar statistical software, problems can arise if the panel data source supplies less than 500 observations on individuals,

which is not an uncommon occurrence. One may have to resort to quadruple precision or generalized inverse routines to conduct such inversions. Regardless of whether one can invert $\widetilde{V}_N$, reliable estimates of its individual elements are unlikely since this matrix contains over 20,000 unique quantities. Problems do not necessarily go away if the sample size were to double to 1000 or quadruple to 2000 observations. Of course, even with these larger samples one would still have little hope of jointly estimating the 10 period model using all over-identifying instrumental variables, since the dimension of $\widetilde{V}_N$ would balloon to 615×615.[18] As we will see in the next section, estimation of sophisticated variants of an ARMA process for errors can readily enlarge the number of equations in ℓ_i even further.

Second, longitudinal surveys sometimes supply very large amounts of data that lead to exceeding memory barriers imposed in software applications, making the implementation of the above MM methods problematic, if not impossible. This issue is especially prevalent in Windows statistical software where the memory barrier currently falls between 1.5 and 2 GB. One must conscientiously accommodate this barrier when using monthly data from a longitudinal survey such as SIPP96 or NLSY79.

There are additional reasons for simplifying estimation within the MM framework, beyond providing options for avoiding the potential computational difficulties described above. Most importantly, less-burdensome methods would offer valuable procedures for carrying out diagnostic tests without the need to estimate the entire model, as the previous discussion assumes. The next section reviews a variety of simplifications in estimation methods.

3.2.5. Estimation with stratified and unbalanced data

Practically all micro data are collected using a stratified sampling frame, meaning that observations with particular characteristics are drawn in proportions differing from those of the true population. Failure to account for this sampling frame in an empirical analysis results in the computation of inconsistent estimates, even when calculating simple statistics such as means. Consequently, naive application of the MM approach produces inconsistent estimates as well.

The question is how to modify MM estimation methods to recognize the implications of the stratified sampling present in longitudinal surveys. These surveys provide weights for use in the calculation of statistics to compensate for nonrandom sampling, but they invariably supply many weights. Besides offering at least one set for each time period for the purpose of computing the appropriate cross section statistics, surveys regularly provide different weights to compensate for over-samples of particular race/ethnic groups

[18] In this simple linear model, approaches exist for reducing the number of instrumental variables in the construction of the ℓ_i's while resulting in little or no loss in estimation efficiency. For example, according to (3.22), near-optimal instrumental variables for the coefficients π, β_1, and β_2 consist of the quantities $\hat{y}_{(t-1)i}$, x_{ti}, and $x_{(t-1)i}$, where $\hat{y}_{(t-1)i}$ represents the fitted value of $y_{(t-1)i}$ regressed on $y_{(t-2)i}, \ldots, y_{-1i}$ and all x_{ki} – with time effects removed from all variables. This yields an ℓ_i with 120 elements.

or low-income families. Which weights should one use in MM procedures when estimating dynamic relationships, and how should these weights be incorporated in forming the ℓ_i's appearing in equation systems (3.8) and (3.19)?

Another important question concerns how to carry out estimation when one has an unbalanced data structure. Whereas balanced samples restrict data to be available for a common set of time periods for each individual included in the analysis, unbalanced samples retain individuals without requiring data for every period. Typically, when faced with unbalanced data, researchers discard observations until they have constructed a balanced sample. The resulting loss of data not only lowers efficiency, but, more fundamentally, it often leads to the selection of nonrepresentative segments of the original sample, and also eliminates a rich source of information for identifying dynamic relationships for sample members seen sporadically or for short horizons.

One might initially surmise that the MM framework can be easily modified to account for unbalanced samples. After all, one can readily portray the ℓ_i's as having a different number of elements – consistent with a different number of time periods – for each individual. However, as demonstrated in Section 6, the formulas reported above do not give the correct representations for the asymptotic distributions of the MM estimators in this case. More sophisticated adjustments are required.

4. Simplified estimation approaches

The following discussion, in conjunction with the next two sections, lays out specific approaches for dealing with the challenges outlined above in a longitudinal data setting. The current section focuses on a variety of simplifications in estimation methods.

The discussion opens with an overview of 3SLS procedures, a well-known special case of the MM framework that yields convenient computational formulas for large systems of equations. It is not straightforward to incorporate predetermined variables in such a procedure, for most 3SLS routines presume common instrumental variables across all equations. The subsequent analysis shows how to surmount this problem with a minimal amount of extra programming and computational burden.

There are considerable advantages to breaking up a longitudinal data estimation problem into parts, allowing researchers to focus on one part of the model at a time. The panel data models introduced in Section 2 provide a rich set of specifications, making the task of choosing among these specifications a formidable endeavor. Not only do they permit flexible parameterizations relating measured variables, but numerous formulations are available for error processes; indeed, far more than can be entertained in standard time series analyses. A researcher rarely knows precisely which parameterizations are consistent with data, and typically must invest considerable effort in performing diagnostic procedures designed to narrow model choices.

This section presents an array of procedures that subdivide the problem of estimating the many parameters introduced in a longitudinal time-series specification into manageable pieces. This multi-step approach permits a researcher to focus on fitting particular

components of the model (such as the AR or MA structure of the error process) without having either to estimate all parameters jointly or to adjust output reported by statistical packages when conducting tests among alternative structures. These procedures offer a powerful set of diagnostic tools useful not only for evaluating the basic features of specifications – such as identifying the orders of ARMA models consistent with the data – but also for discovering reliable values for parameters that can serve as starting values for the larger estimation exercises.

4.1. Several important computational simplifications

Three approaches assist in dealing with the computational challenges of MM methods outlined in Section 3.2. The first relates to the design of the estimation problem so as to permit application of multi-step procedures requiring computation of only subsets of the parameters at a time. The second involves specialization of MM methods to 3SLS procedures, irrespective of whether estimation is linear or nonlinear. Finally, the third proposes adaptations of 3SLS procedures to incorporate predetermined variables as instrumental variables in estimation. This subsection elaborates on each of these computational simplifications.

4.1.1. A condition simplifying multi-step estimation

In the application of estimation procedures considered in Section 4.2, it is very convenient to limit the number of parameters estimated at any stage of the analysis by fixing a subset of the parameters at a consistent estimate obtained from a previous stage. With $f_i(\gamma_0, \mu_0)$ replacing $f_i(\gamma_0)$ in model (3.7), these estimation procedures can be described as computing an estimate $\tilde{\gamma}$ for γ_0 by minimizing (3.10) with $f_i(\gamma, \hat{\mu})$ substituted for $f_i(\gamma)$, where $\hat{\mu}$ is a consistent estimate of μ_0. The application of NIV produces standard errors and test statistics for $\tilde{\gamma}$ according to (3.11), with gradient matrix $\frac{\partial f_i}{\partial \gamma'}|_{\tilde{\gamma},\hat{\mu}}$ replacing the gradient matrix $\frac{\partial f_i}{\partial \gamma'}|_{\tilde{\gamma}}$ in the expression for the asymptotic variance–covariance matrix. In general, these standard errors and test statistics are invalid because they ignore any correction for estimation error induced by imperfect knowledge of μ_0. However, given a special set of conditions stated in the following proposition, no correction for estimation error is needed when computing standard errors.

PROPOSITION 4.1. *Suppose $f_i(\gamma_0, \mu_0)$ replaces $f_i(\gamma_0)$ in model (3.7) and this vector of structural equations satisfies the property*

$$E\left(\left.\frac{\partial f_i}{\partial \mu'}\right|_{\gamma_0,\mu_0}\right) = 0. \tag{4.1}$$

Then, NIV applied to the system of equations $f_i(\gamma, \hat{\mu})$, where $\hat{\mu}$ is a consistent estimate of μ_0, yields an estimator $\tilde{\gamma}$ whose asymptotic distribution is given by (3.11) with the

gradient matrix $\frac{\partial f_i(\gamma,\hat{\mu})}{\partial \gamma'}|_{\tilde{\gamma}}$ *replacing the matrix* $\frac{\partial f_i}{\partial \gamma'}|_{\hat{\gamma}}$ *in the formula for the variance–covariance matrix.*[19]

This proposition has two important implications for the following analysis: (i) the application of a standard NIV procedure for these cases not only produces a consistent estimate of γ_0 when $\hat{\mu}$ is treated as fixed, but also reports asymptotically valid standard errors and test statistics; and (ii) joint estimation of γ_0 and μ_0 by NIV using the equations in f_i will not lead to an improvement in asymptotic efficiency. Many of the econometric specifications considered in the subsequent discussion satisfy this condition, so attention can be focused on estimating and testing hypotheses involving subsets of the parameters.

4.1.2. Three-stage least squares

Conventional 3SLS analysis – a special case of NIV – maintains the assumption that the error vector ξ_i is homoscedastic or distributed independently of Q_i, where the vector Q_i contains all the linearly independent elements of the Q_{ti}'s. Such an assumption is often satisfied in longitudinal analyses. Thus, the variance–covariance matrix $E\{\xi\xi'\} = E\{\xi_i\xi_i'|Q_i\}$ is constant across observations.

Accordingly, in the formula for the NIV distance function given by (3.10), one can replace the estimated matrices $\tilde{\xi}_i\tilde{\xi}_i'$ by a consistent estimate of the variance–covariance

[19] To demonstrate this proposition, define $F(\gamma, \hat{\mu})$ as the function given by (3.10), with $f_i(\gamma, \hat{\mu})$ substituted for $f_i(\gamma)$. Minimizing F defines the estimator $\tilde{\gamma}$ by the first-order condition

$$\overset{\circ}{F}(\tilde{\gamma}, \hat{\mu}) \equiv \left.\frac{\partial F}{\partial \gamma}\right|_{\tilde{\gamma},\hat{\mu}} = 0.$$

Taking an exact first-order Taylor expansion of this system of equations in γ and μ around the true values of these parameters yields

$$\overset{\circ}{F}(\gamma_0, \mu_0) + \overset{\circ\circ}{F}_{\gamma}(\gamma^*, \mu^*)(\tilde{\gamma} - \gamma_0) + \overset{\circ\circ}{F}_{\mu}(\gamma^*, \mu^*)(\hat{\mu} - \mu_0) = 0,$$

where

$$\overset{\circ\circ}{F}_{\gamma} = \frac{\partial \overset{\circ}{F}}{\partial \gamma'}, \qquad \overset{\circ\circ}{F}_{\mu} = \frac{\partial \overset{\circ}{F}}{\partial \mu'}$$

and the values (γ^*, μ^*) lie between $(\tilde{\gamma}, \hat{\mu})$ and (γ_0, μ_0). Solving these equations for $\tilde{\gamma} - \gamma_0$, it can be shown using the standard regularity assumptions that

$$\operatorname{plim}\{\sqrt{N}(\tilde{\gamma} - \gamma_0)\} = \operatorname{plim}\{G[\overset{\circ}{F}(\gamma_0, \mu_0)/\sqrt{N} + H\sqrt{N}(\hat{\mu} - \mu_0)]\},$$

where

$$G^{-1} = \operatorname{plim}\{-\overset{\circ\circ}{F}_{\gamma}(\gamma_0, \mu_0)/N\} \quad \text{and} \quad H = \operatorname{plim}\{\overset{\circ\circ}{F}_{\mu}(\gamma_0, \mu_0)/N\}.$$

Condition (4.1) implies that $H = 0$. Consequently, $\sqrt{N}(\tilde{\gamma} - \gamma_0)$ has the same asymptotic distribution as $G\overset{\circ}{F}(\gamma_0, \mu_0)/\sqrt{N}$; and (3.11) gives a large-sample approximation to the distribution of $\gamma_0 + G\overset{\circ}{F}(\gamma_0, \mu_0)/N$.

matrix of the error vector ξ. Designating this estimate as $\widetilde{E}(\xi\xi')$, a standard calculation for this quantity is

$$\widetilde{E}(\xi\xi') = \frac{1}{N}\sum_{i=1}^{N}\tilde{\xi}_i\tilde{\xi}_i'. \tag{4.2}$$

One computes the 3SLS estimator, then, by minimizing the distance function

$$\left[\Sigma_i f_i' \Delta_i'\right]\left[\frac{1}{N}\Sigma_i\left(\Delta_i \widetilde{E}(\xi\xi')\Delta_i'\right)\right]^{-1}\left[\Sigma_i \Delta_i f_i\right]. \tag{4.3}$$

The asymptotic normal distribution for the 3SLS estimator is

$$\hat{\gamma} \overset{\cdot}{\sim} N\left(\gamma_0, \left[\left[\Sigma_i \frac{\partial f_i'}{\partial \gamma}\bigg|_{\hat{\gamma}} \Delta_i'\right]\left[\Sigma_i\left(\Delta_i \widetilde{E}(\xi\xi')\Delta_i'\right)\right]^{-1}\left[\Sigma_i \Delta_i \frac{\partial f_i}{\partial \gamma'}\bigg|_{\hat{\gamma}}\right]\right]^{-1}\right). \tag{4.4}$$

In those software packages specifying a common set of instrumental variables for each equation, the nonlinear 3SLS estimator $\hat{\gamma}$ is defined by that value of γ minimizing the function

$$\left[\Sigma_i f_i(\gamma)' \otimes Q_i'\right]\left[\widetilde{E}(\xi\xi') \otimes \frac{1}{N}\Sigma_i Q_i Q_i'\right]^{-1}\left[\Sigma_i f_i(\gamma) \otimes Q_i\right]. \tag{4.5}$$

The matrix $[\widetilde{E}(\xi\xi') \otimes \frac{1}{N}\Sigma_i Q_i Q_i']$ in these expressions corresponds to the matrix $\widetilde{V}_N$ appearing in the MM distance function (3.4). Even for large equation systems including many measured variables and time periods, this construction of $\widetilde{V}_N$ is easily computed, as is its inverse. The dimension of the matrix $\widetilde{E}(\xi\xi')$ is merely the number of structural equations (or the number of time periods if there exists a single equation per period), whereas the dimension of $\frac{1}{N}\Sigma_i Q_i Q_i'$ corresponds to the total number of instrumental variables used in the analysis.[20] This estimator is consistent for γ_0, and in large samples $\hat{\gamma}$ approximately follows a multivariate normal distribution given by

$$\hat{\gamma} \overset{\cdot}{\sim} N\left(\gamma_0, \left[\left[\Sigma_i \frac{\partial f_i'}{\partial \gamma}\bigg|_{\hat{\gamma}} \otimes Q_i'\right]\left[\widetilde{E}(\xi\xi') \otimes \Sigma_i Q_i Q_i'\right]^{-1}\left[\Sigma_i \frac{\partial f_i}{\partial \gamma'}\bigg|_{\hat{\gamma}} \otimes Q_i\right]\right]^{-1}\right). \tag{4.6}$$

4.1.3. Adding equations to account for predetermined variables

Conventional linear/nonlinear 3SLS computer programs (including seemingly unrelated regression routines) do not permit inclusion of predetermined variables as instrumental variables. In conventional 3SLS programs, one must classify a variable either as

[20] If a linear/nonlinear 3SLS procedure encounters difficulties in inverting the matrix $\frac{1}{N}\Sigma_i Q_i Q_i'$, it eliminates elements of Q_i until this matrix becomes invertible. This results in no effective loss of efficiency since this smaller variant of Q_i spans the same space as the original instrumental variable vector; consequently, it does as well in predicting all endogenous components of the structural equation.

endogenous or as an instrumental variable for the entire system of equations. Including predetermined variables in the list of instruments or in any prediction equation for endogenous variables in the application of these programs will result in inconsistent parameter estimates.

Fortunately, one can devise a relatively simple and computationally feasible method for using predetermined variables as instruments in the estimation of model (3.18) within a standard 3SLS program. This approach adds several new structural equations to the model. Suppose that the predetermined variables can be related to the exogenous variables by the regression equations

$$P_{ti} = \delta_t Q_i + \eta_{ti}, \quad t = 1, \ldots, T, \tag{4.7}$$

where δ_t is a matrix of coefficients, and the errors η_{ti} are distributed independently across observations and are independent of the exogenous variables X_i (and, therefore, of the elements of the instrumental variables Q_i).

For the moment assume that the η_{ti} are observed (i.e., that data are provided for these errors) and that g_{ti} in (3.18) embodies a single structural equation in period t. To model (3.18), add the structural equations $\eta_{ti} g_{ti} = \xi_{ti}$ for $t = 1, \ldots, T$. This creates an expanded system of equations that can be compactly expressed in terms of model (3.7) with

$$f_{ti} = \begin{pmatrix} g_{ti} \\ \eta_{ti} g_{ti} \end{pmatrix}. \tag{4.8}$$

The error vector ξ_i implied by this specification has a zero mean and is homoscedastic across individuals given the assumption of the independence of η_{ti} with g_{ti} and Q_i.

Thus, this particular specification of model (3.7) can be estimated by a standard 3SLS procedure with a common set of instrumental variables Q_i used for all equations. The estimator calculated from this procedure is consistent and the output reported by this computation is asymptotically valid. Moreover, its asymptotic efficiency at least matches that of the NIV estimator computed with ℓ_i specified by (3.19), which uses all the predetermined variables as instrumental variables in the estimation of γ.[21]

This 3SLS estimator can be easily modified to account for the fact that the disturbances η_{ti} are not directly observed. In particular, in the specification of the f_{ti}'s given by (4.8), one merely needs to replace the η_{ti}'s by their corresponding LS or GLS residuals – defined by $\hat{\eta}_{ti} = P_{ti} - \hat{\delta}_t Q_i$. Application of the standard 3SLS program to this modified specification of (4.8) continues to produce asymptotically valid standard errors and test statistics, and an estimator with the same large sample properties as one computed using the true η_{ti}'s. This conclusion follows directly from Proposition 4.1 by interpreting μ as $(\delta_1', \ldots, \delta_T')'$.

[21] This claim follows from the observation that the quantity $[g_{ti} \otimes (Q_{ti}', P_{ti}')']$ – used in (3.19) to compute the instrumental variable estimator – is a strict linear combination of the quantity $[(g_{ti} \otimes (1, \eta_{ti}')') \otimes Q_i]$ – used in the computation of 3SLS with Q_i used as instrumental variables for all equations. This observation presumes that Q_i contains a constant.

The following proposition summarizes the key results that allow one to compute an estimate of the coefficients γ_0 appearing in model (3.7) using both the available exogenous and predetermined variables as instrumental variables.

PROPOSITION 4.2. *With Q_i as the instrumental variables, 3SLS applied to model (3.7) with the specification*

$$f_{ti}(\gamma, \hat{\mu}) = \begin{pmatrix} g_{ti} \\ g_{ti} \otimes \hat{\eta}_{ti} \end{pmatrix}, \quad \text{for } t = 1, \ldots, T \tag{4.9}$$

yields an estimator for γ_0 whose large-sample distribution is given by (4.6) and whose asymptotic efficiency attains that associated with the instrumental variable estimator obtained by minimizing (4.3) with ℓ_i specified by (3.19).

This proposition implies two important results: (i) standard 3SLS estimation of this model produces standard errors and test statistics that are asymptotically valid; and (ii) the estimate of γ_0 computed by this procedure is as efficient as one calculated using the exogenous variables as instruments for all equations and the predetermined variables as instruments for the subset of equations for which they are appropriate.

4.1.4. Incorporating optimal instruments with predetermined variables in 3SLS

If one desires to exploit a near-optimal set of instrumental variables in the application of 3SLS estimation, analogous to the set characterized in Section 3.2.2 for the general MM case, a straightforward modification of Equations (4.9) achieves this formulation. One must first transform equation system (3.18) into the form given by model (3.20). Then, $g_{ti}^{\#}$ replaces g_{ti} in (4.9).

The formulation of the $\hat{\eta}_{ti}$'s appearing in (4.9) changes as well. According to (3.22), the near-optimal instrumental variables for the system are the $\widehat{Q}_{ti}^{\#}$'s representing the fitted values obtained by regressing the quantities $\frac{\partial g_{ti}^{\#}}{\partial \gamma_t}|_{\tilde{\gamma}}$ on functions of the elements of Q_i and P_{ti}. Replacing P_{ti} by $\widehat{Q}_{ti}^{\#}$ in regression equation (4.7) produces the residuals $\hat{\eta}_{ti} = \widehat{Q}_{ti}^{\#} - \hat{\delta}_t Q_i$. It is the linearly independent components of these residuals that go into constructing the expanded system of structural equations given by (4.9).

Simpler procedures exist for attaining most of the gains achievable through explicit use of optimal instrumental variables in 3SLS. If one specifies a Q_i and P_{ti}'s in the original formulation of 3SLS that virtually span the space of $\widehat{Q}_{ti}^{\#}$, then few efficiency improvements are possible with the actual use of $\widehat{Q}_{ti}^{\#}$.

4.2. Estimating subsets of parameters

Proposition 4.1 serves as the cornerstone for many estimation methods that rely on multi-stage procedures wherein a later stage conditions on parameter values estimated in earlier stages without recognizing any estimation error associated with the fixed parameters. Generalized least squares represents the classic example of such a procedure. A first

stage estimates parameters describing the error structure, and a second stage uses this estimated structure to compute a weighting matrix in the application of least squares. The second stage ignores that the weighting matrix depends on estimated parameters. The form of this estimation method satisfies the conditions listed in Proposition 4.1.

The following discussion exploits this proposition to provide a variety of procedures allowing researchers to subdivide the problem of estimating parameters of sophisticated longitudinal specifications into a multi-stage approach. In each step, the application of familiar estimation routines yield valid test statistics that are useful for discovering which parts of a model fit the data without having to specify all parts together. The analysis assumes that a panel data source offers a fixed number of time periods T and asymptotic results depend on a large number of individuals N. Later sections develop results when T differs across individuals.

4.2.1. Distinguishing the different parameter subsets

The parameters of the longitudinal specifications introduced in Section 2 may be grouped into two sets: the first – hereafter called the structural coefficients – consists of those coefficients included in the matrices Π, Ψ, and B which relate measured variables and determine distributed lag relationships; and the second set – hereafter called the covariance parameters and denoted by the vector α – consists of those parameters involved in the specification of the covariance matrix $\Theta \equiv E\{U_i U_i'\}$. If a researcher considers a specification other than a DSEM of the sort described in Section 2.3, then "structural parameters" refer to those coefficients appearing in the specification of f_{ti} characterizing the dynamic relationships linking Y_{ti}, Z_{ti}, and X_{ti}. The covariance parameters α include coefficients of the lag polynomials $a(L)$ and $m(L)$, variances of white noise and any permanent components if they are present, and the elements of the covariance matrix summarizing information on initial conditions.[22]

4.2.2. Estimation of structural coefficients

If the sole aim of a longitudinal analysis is to estimate the coefficients of distributed lag structures relating measured variables, it is well known that one can carry out this

[22] The following analysis does not present any formal identification conditions. For the standard multivariate ARMA model, Kashyap and Nasburg (1974) develop necessary and sufficient conditions for identification and Hannan (1969) presents sufficient conditions. These conditions are not easily applied in practice and panel data introduce additional complications. The length of the time series, for example, becomes a crucial factor. Also, the treatment of initial conditions reduces the effective length of the panel and at the same time introduces new parameters. Considering an error specification that combines a permanent component with a time series process does not complicate the identification problem. First differencing equations eliminates permanent components, and it does not introduce any new parameters. Thus, introducing a permanent component only has the effect of reducing the length of the time series by one period, and identification will be lost only in those cases in which the orders of the autoregressive and the moving average components are sufficiently high to make the length of the time series a crucial factor.

estimation without assuming anything about the stochastic process generating the disturbances U_{ti}. Given a large sample of individuals, the variance–covariance matrix Θ can be left unconstrained and estimated by standard methods using residuals computed for U_{ti}.

When one disregards predetermined variables as instruments, standard procedures can be directly applied to estimate the coefficients of model (2.10) and to test hypotheses regarding these coefficients. Joint generalized least squares can be employed to estimate this equation system and the parameter vector β when no predetermined or endogenous variables appear in this system (i.e., when $\pi = 0$ and $\psi = 0$). Otherwise, 3SLS can be applied to estimate π, ψ and β in (2.10), with X_i used as instrumental variables and with Z_i treated as endogenous. To account for the restrictions implied by distributed lag structures, one must impose equality constraints across equations when applying these estimation methods.

Expressed in terms of the simultaneous equation framework outlined in Section 3, define the vector f_i as the system of equations in model (2.10), which is given by

$$f_i^{(1)}(\pi, \psi, \beta) = y_i - Y_i\pi - Z_i\psi - X_i\beta = U_i. \tag{4.10}$$

Applying 3SLS (or joint generalized least squares) to model (2.10) amounts to specifying $f_i = f_i^{(1)}$ in (3.7) and computing an estimate of $\gamma = (\pi', \psi', \beta')$ by minimizing (3.10) or (4.5), including in the instrumental variables Q_i all the unique elements of X_i and a constant.

When one wants to use predetermined variables as instruments in the estimation of the structural coefficients, then $f_i^{(1)}$ is modified according to Proposition 4.2, with $g_i = y_i - Y_i\pi - Z_i\psi - X_i\beta$ and $\gamma = (\pi', \psi', \beta')$.[23] Instrumental variable estimation is then carried out as outlined in this proposition, with Q_i consisting of all the unique elements of X_i along with a constant. In the subsequent discussion all references to (4.10) as the specification of $f_i^{(1)}$ are meant to imply this modified formulation of $f_i^{(1)}$ in which predetermined variables are also exploited as instruments in the estimation of model (2.10).

These estimation methods offer a simple framework for performing preliminary data analysis to determine the order of the polynomials associated with distributed lags (i.e., $\Pi(L)$, $\Psi(L)$ and $B(L)$ in (2.8)) and to test whether the coefficients of these polynomials are constant across periods. This form of data analysis involves standard tests of linear hypotheses.

While specification (2.9) presumes that a researcher wishes to analyze only a single structural equation per period, it is straightforward to modify this specification to permit analysis of a multicolumn vector, and of the coefficients Π_j, Ψ_j and B_j as matrices of parameters.

[23] The past values of Y_t will not be predetermined, except for special cases of the stochastic process generating U_{ti}. Thus, in applications where a researcher wants to estimate structural coefficients in complete ignorance of the variance–covariance matrix Θ, lagged values of Y_t cannot be considered as predetermined variables.

4.3. Estimation of covariance parameters

As briefly noted in Section 2.4, estimating parameters determining the autocorrelation structure of the disturbances U_{ti} is far more difficult than estimating structural coefficients linking measured variables. Part of the reason for difficulties arises from the fact that the U_{ti}'s are unobserved and must themselves be estimated. Further complications come about with the presence of moving-average error structures that (i) introduce problems due to initial conditions and (ii) require sophisticated transformations to isolate ARMA coefficients.

Two basic approaches provide for the estimation of coefficients determining the specification of the variance–covariance matrix Θ for the error vector U_i appearing in (4.10). One expresses the elements of Θ in terms of the underlying ARMA coefficients and estimates these coefficients using fitted values of U_i and nonlinear regression methods, with adjustments introduced to compute standard errors and test statistics that account for use of estimated values in place of the true values of U_i. The second set of approaches further subdivides the estimation problem by allowing researchers to estimate autoregressive and moving-average coefficients separately. These latter procedures are especially applicable when the particular orders of the AR and MA lags in the ARMA model are unknown and one needs procedures for testing and identifying the basic form of the lag structure. Whereas this subsection summarizes features of the first approach, Sections 4.4–4.7 outline estimation procedures relevant for the second approach.

4.3.1. Framework for estimating variance and covariance parameters

Suppose for the moment that one directly observes U_i. Consider the problem of estimating a single element of the covariance matrix Θ, say the one corresponding to the covariance $E\{U_{ti}U_{(t-k)i}\}$, denoted by the parameter θ_{kt}. A simple way to proceed for obtaining an estimate of θ_{kt} is to consider the regression equation

$$U_{ti}U_{(t-k)i} = \theta_{kt} + V_{kti} \tag{4.11}$$

for $i = 1, \ldots, N$, where $U_{ti}U_{(t-k)i}$ is the dependent variable and V_{kti} is an error term defined to have a zero mean. Since the dependent variables and, thus, the V_{kti} are independently distributed across individuals, it is evident that least squares estimation of Equation (4.11) using cross-sectional data on individuals will yield a consistent estimate for θ_{kt} and valid test statistics.

Combining these regression equations for estimating the different elements of the covariance matrix Θ into a single seemingly unrelated regression model provides a convenient framework for estimating the set of covariance parameters. In particular, stacking Equations (4.11) for various values of t and k (i.e., $t = 1, \ldots, T$, and

$k = 0, \ldots, t-1$) for a given individual i yields the system of equations

$$\begin{bmatrix} U_{Ti}U_{Ti} \\ U_{Ti}U_{(T-1)i} \\ \vdots \\ U_{Ti}U_{1i} \\ U_{(T-1)i}U_{(T-1)i} \\ \vdots \\ U_{(T-1)i}U_{1i} \\ U_{(T-2)i}U_{(T-2)i} \\ \vdots \\ U_{1i}U_{1i} \end{bmatrix} \equiv \mathrm{St}\{U_i U_i'\} = \theta + V_i, \tag{4.12}$$

where $\mathrm{St}\{\cdot\}$ denotes an operator that stacks the transposes of the rows of a matrix after deleting all elements that lie below the diagonal, $\theta \equiv \mathrm{St}\{\Theta\}$ is a vector of intercepts, and V_i is an error vector composed of the disturbances V_{kti} of Equation (4.11) for the implied values of t and k. Writing the intercepts of model (4.12) as functions of the form $\theta = \mathrm{St}\{\Theta(\alpha)\} \equiv \theta(\alpha)$ and given data for U_i, one can – using data on individuals – compute an estimate for α and test hypotheses concerning its structure using a conventional nonlinear joint generalized least squares procedure.

One encounters three formidable challenges in utilizing equation system (4.12) to estimate parameters of the ARMA process (i.e., the coefficients a, m and Σ in (2.3) and (2.4)) determining the elements of θ. First, one requires a complete specification for Θ relating how each element of θ depends on coefficients a, m and Σ. As illustrated below, such an exercise is not as easy as one may initially surmise. Second, an obvious challenge involves unavailability of data on U_i. Using residuals $\widehat{U}_i$ in place of the actual U_i typically implies that standard errors and test statistics must be adjusted to account for estimation error present in dependent variables. Finally, for even short panels, the number of equations in system (4.12) is quite large, making estimation burdensome as outlined in Section 3.2. The subsequent discussion deals with each of these problems.

4.3.2. Specification of variance–covariance matrix accounting for initial conditions

Regarding the first challenge, any development of a specification for Θ requires a complete understanding of the initial conditions problem associated with ARMA processes. The consequence of assuming that the disturbances appearing in (2.8) follow the error specification given by (2.3) is the imposition of restrictions on the variance–covariance matrix $\Theta \equiv E\{U_i U_i'\}$ associated with the stacked representation of the DSEM given by (2.10). The following analysis describes the exact restrictions on autocovariances implied by this error specification, and formulates an explicit parameterization for Θ. Appendix A expands upon this discussion and presents a more elaborate derivation of this parameterization. While solutions to this problem can be found in the panel data

literature for pure autoregressive or pure moving-average schemes, none are available for mixed ARMA processes.

According to (2.3) with AR and MA coefficients constant over time, U_i is determined by the system of equations

$$U_i = \begin{pmatrix} U_{Ti} \\ \vdots \\ U_{1i} \end{pmatrix} = - \begin{pmatrix} \sum_{j=1}^{p} a_j U_{(T-j)i} \\ \vdots \\ \sum_{j=1}^{p} a_j U_{(1-j)i} \end{pmatrix} + \begin{pmatrix} \sum_{j=0}^{q} m_j \varepsilon_{(T-j)i} \\ \vdots \\ \sum_{j=0}^{q} m_j \varepsilon_{(1-j)i} \end{pmatrix}. \tag{4.13}$$

This system does not represent a one-to-one transformation from the ε_{ti}'s, $t = 1, \ldots, T$, to U_i. One cannot derive the covariance matrix for U_i from (4.13) given only the distributional assumptions for $\varepsilon_{Ti}, \ldots, \varepsilon_{1i}$. Also appearing in (4.13) are the variables $U_{0i}, \ldots, U_{(1-p)i}$, and $\varepsilon_{0i}, \ldots, \varepsilon_{(1-q)i}$ which are known in the time series literature as initial conditions or starting values for the error process. To derive a parameterization for Θ, one requires a specification of initial conditions.

Conventional time series techniques that consider starting values as known constants (usually chosen to be zero) result in inconsistent estimates for the parameters of the error process if the technique is applied in a panel data analysis where T is fixed. Similarly, time series procedures that "backforecast", or treat initial conditions as parameters, introduce an incidental parameters problem in a panel data analysis which, under most circumstances, also leads to inconsistent estimates for all parameters of the error process.[24] A third way to deal with these initial conditions for the disturbances is to treat them as random variables. This is the procedure followed below. Treating initial conditions as random variables avoids problems with inconsistency by introducing only a finite number of new parameters: those determining the distribution of the starting values and those relating the distribution of the starting values to the distribution of disturbances realized in periods 1 through T.

There are several complications associated with choosing a distribution for the initial conditions specified above. If we assume that the stochastic process generating disturbances during the sample period is also operative prior to this period, then one would expect the U_{ki}'s, $k = (1-p), \ldots, 0$, to be not only correlated with one another and with the ε_{ti}'s, $t = (1-q), \ldots, 0$, but also with all the U_{ti}'s realized after period 0. Furthermore, the correlations relating these variables will, in general, depend directly on parameters of the ARMA process given by (2.3), and one must account for these restrictions to achieve efficient estimation.

A natural approach for dealing with this specification of the correlation properties of initial conditions is to assume that the time series process generating disturbances over the sample period started some time prior to this period at an unknown date and under an unknown set of circumstances. In particular, assume the ARMA process given by (2.3) started in the *finite* past between periods ℓ_2 and ℓ_1 with $\ell_2 < \ell_1$ and with ℓ_1 occurring at least $p - q + 1$ periods prior to the first sample observation. One can write

[24] See Hsiao (1986) for further discussion of this problem.

each of the U_{ti}'s realized after period ℓ_1 as a moving-average scheme of the form

$$U_{ti} = \sum_{j=0}^{t-\ell_1} \zeta_j \varepsilon_{(t-j)i} + \sum_{j=t-\ell_1+1}^{t-\ell_2} \zeta_j \phi_{(t-j)i}, \tag{4.14}$$

where the ζ_j coefficients are defined as $\zeta_0 \equiv 1$ and $\zeta_j \equiv m_j - \sum_{h=1}^{j} a_h \zeta_{j-h}$ for $j > 0$; the ε_{ki}'s for $k \geqslant \ell_1$ are white noise; and the ϕ_{ki}'s are error terms distributed independently of $\varepsilon_{\tau i}$ for $\tau \geqslant \ell_1$. Formally, one can derive a relation like (4.14) by starting with the ARMA representation $U_{ti} = -\sum_{j=1}^{p} a_j U_{(t-j)i} + \sum_{j=0}^{q} m_j \varepsilon_{(t-j)i}$ and successively substituting out for past $U_{(t-j)i}$'s using their ARMA representations until $t - j = \ell_1 - 1$. The ϕ_{ki}'s in (4.14) may be interpreted as the true starting values of the ARMA process. Specifying the distribution of these variables determines exactly how and when the ARMA process generating the U_{ti}'s began.

Given this moving-average representation of the ARMA process generating transitory components, one can derive a complete specification for the covariance matrix of U_i. A change of variables simplifies the derivation. Define the random variable e_{ti} by $e_{ti} = \sum_{j=0}^{p} a_j U_{(t-j)i}$ and consider a linear transformation from the vector $U_i' = (U_{Ti}, \ldots, U_{1i})$ to a new vector $(e_i', U_{(1)i}')$, where $e_i' = (e_{Ti}, \ldots, e_{(p+1)i})$ and $U_{(1)i}' = (U_{pi}, \ldots, U_{1i})$. With F denoting this linear transformation,

$$\begin{pmatrix} e_i \\ U_{(1)i} \end{pmatrix} = F U_i, \tag{4.15}$$

where F is a $T \times T$ matrix that can be partitioned as

$$F = \begin{pmatrix} A_{(T-p)\times(T)} \\ O_{p\times(T-p)} \quad I_{p\times p} \end{pmatrix},$$

with A representing a diagonal band matrix with the elements $(a_0, \ldots, a_p)$ running down the diagonal.[25] The following analysis first develops the covariance matrix for the vector $(e_i', U_{(1)i}')'$, and then uses this result to obtain Θ. The covariance matrix of $(e_i', U_{(1)i}')'$ is conveniently partitioned into three blocks: the (1, 1) block, $E\{e_i e_i'\}$; the (2, 2) block, $E\{U_{(1)i} U_{(1)i}'\}$; and the (2, 1) block $E\{U_{(1)i} e_i'\}$. Below we derive the implied parameterization for each of these blocks in turn.

[25] A matrix Q is a diagonal band with the elements (a, b, c, d) running down the diagonal if it has the form

$$Q = \begin{bmatrix} a & b & c & d & & & & & & & \\ & a & b & c & d & & & & & & \\ & & & & & \cdots & & & & & \\ 0 & & & & & & \cdots & & & 0 & \\ & & & & & & & \cdots & & & \\ & & & & & & & & a & b & c & d \\ & & & & & & & & & a & b & c & d \end{bmatrix}.$$

Since $e_{ti} = a(L)U_{ti} = m(L)\varepsilon_{ti}$, we see that the e_{ti}'s, for $t = (p+1), \ldots, T$, are generated by a pure moving-average process. Their covariance matrix, then, is determined uniquely by the relationships

$$E\{e_{ti}e_{(t-h)i}\} = \begin{cases} \sum_{j=0}^{q-h} m_{j+h}E[\varepsilon^2_{(t-h-j)i}]m_j = \sum_{j=0}^{q-h} m_{j+h}\sigma^2 m_j \\ \quad \text{for } 0 \leqslant h \leqslant q, \\ 0 \quad \text{for } h > q, \end{cases} \tag{4.16}$$

for $t = p+1, \ldots, T$ and $p+1 \leqslant t-h \leqslant t$.

The moving-average expression for the U_{ti}'s, $t = 1, \ldots, p$, given by (4.14) provides the only information available for determining a parameterization for $E\{U_{(1)i}U'_{(1)i}\}$. Inspection of this expression reveals that the elements of $U_{(1)i}$ depend directly on the random variables ϕ_{ki} and on the number of periods since these variables were realized. Unless one is willing to be very specific about how and when the ARMA process generating transitory components started for each individual in the sample, nothing can be said about the number or the correlation properties of the ϕ_{ki}'s, or about how far in the past they were realized. To avoid specifying this information, one can simply assume that starting times and the ϕ_{ki}'s are randomly distributed over the population, in which case no restrictions are implied for the covariance structure associated with $U_{(1)i}$. Thus, assume

$$E\big(U_{(1)i}U'_{(1)i}\big) = \Upsilon, \tag{4.17}$$

where Υ is an unconstrained, positive definite, symmetric matrix. As a consequence of this assumption, the time series process generating disturbances need not be stationary.

Finally, one requires a specification for $E\{U_{(1)i}e'_i\}$. Using the moving-average representation for e_{ti}, $t = (p+1), \ldots, T$, and those for U_{ki}, $k = 1, \ldots, p$, given by (4.14), the implied covariance between e_{ti} and $U_{(t-h)i}$ is

$$E\{e_{ti}U_{(t-h)i}\} = \begin{cases} \sum_{j=0}^{q-h} m_{j+h}E[\varepsilon^2_{(t-h-j)i}]\zeta_j = \sum_{j=0}^{q-h} m_{j+h}\sigma^2\zeta_j \\ \quad \text{for } 0 < h \leqslant q, \\ 0 \quad \text{for } h > q. \end{cases} \tag{4.18}$$

Implicit in this formula is the nonrestrictive assumption that the ARMA process for disturbances starts prior to period $p-q+1$,[26] and, as a consequence, covariance terms like $E\{e_{ki}\phi_{si}\}$ do not appear. An attractive feature of this formulation for the covariance is that no new parameters appear in the expression. If one is willing to introduce new parameters into the analysis, it is possible to avoid constructing the ζ_j's and imposing the restrictions implied by (4.18). In particular, one can simply treat the nonzero values of $E\{e_{ti}U_{(t-h)i}\}$ as arbitrary parameters and estimate them directly along with

[26] This assumption concerning the starting time of the ARMA process generating the U_{ti}'s follows immediately from the assumption that $U_{(p+1)i}$ can be represented by the specification given by (2.3). This restriction ensures that no ϕ_{ki}'s appear in the moving average component of (2.3) for $t = p+1$.

the other parameters of the error process. While this alternative parameterization avoids the need for imposing some nonlinear restrictions, it has the disadvantage of reducing the efficiency of estimation; and in some instances, it can destroy the identification of some parameters of the error process if the time series supplied by the panel dataset is short.

The relations given by (4.16)–(4.18) imply an explicit parameterization for the covariance matrix associated with the vector $(e_i', U_{(1)i}')'$. Denote this matrix by Ω. Since this vector and U_i are linearly related according to Equation (4.15), it follows that

$$E\{U_i U_i'\} = F^{-1}\Omega F^{-1\prime}. \tag{4.19}$$

This parameterization imposes all of the covariance restrictions implied by the ARMA process, unless one is willing to introduce precise information about how and when this process began. Appendix A presents explicit expressions for θ that impose all nonlinear constraints. These formulas have relatively simple representations, thus making them particularly useful when applying estimation procedures.

The above treatment of initial conditions induces a source of nonstationarity in the U_{ti}'s, even when all the coefficients of the ARMA model and the variances of white noise are constant over time. Permitting the AR coefficients to be different over time changes the form of the matrix A in a straightforward way, and allowing the MA coefficients to differ alters the form of the matrix Ω.[27] In conventional time-series analyses, these generalizations are either not possible or introduce substantial complications in estimation.

4.3.3. Joint estimation of structural coefficients and covariance parameters

With a specification of θ in hand, we now turn to the second challenge, which involves estimation of the covariance parameters without direct data on the U_i's in (4.12). With observations on U_i unavailable, combining equation systems (4.10) and (4.12) to estimate structural coefficients and covariance parameters jointly offers a conceptually simple framework for estimating the elements of θ or α.

To describe this estimation approach in terms of the nonlinear system of simultaneous equations given by (3.7), split the vector of structural equations f_i in (3.7) into two subvectors $f_i = (f_i^{(1)\prime}, f_i^{(2)\prime})'$. The system of equations $f_i^{(1)}$ given by (4.10) is used in 3SLS estimation of the structural coefficients. In the specification of model (4.12), substitute $y_i - Y_i\pi - Z_i\psi - X_i\beta$ for U_i to obtain the vector of equations

$$\begin{aligned} f_i^{(2)}\big(\pi, \psi, \beta, \theta(\alpha)\big) = \mathrm{St}\big\{&(y_i - Y_i\pi - Z_i\psi - X_i\beta) \\ &\times (y_i - Y_i\pi - Z_i\psi - X_i\beta)'\big\} - \theta(\alpha) = V_i. \end{aligned} \tag{4.20}$$

[27] Baltagi (2002) surveys recent developments of panel data methods for estimating parameters in the presence of several varieties of nonstationarity popular in the times series literature.

Combining (4.10) and (4.20) to form $f_i(\pi, \psi, \beta, \theta(\alpha)) = (f_i^{(1)\prime}, f_i^{(2)\prime})'$ satisfies the assumptions of model (3.7).[28] Since U_i is assumed to be independent of X_i, these exogenous variables constitute valid instruments for all the equations incorporated in f_i. Thus, 3SLS applied to model (3.7) with this specification of f_i produces consistent estimates of π, ψ, β, and θ (or α) and asymptotically valid standard errors and test statistics.[29]

Simultaneously estimating structural coefficients and covariance parameters yields estimates that are, in general, more efficient than those obtained from the other methods outlined in this paper. There are two sources for this increase in efficiency. First, in those instances in which the third moments of U_i are nonzero (which implies that $E\{U_i V_i'\} \neq 0$), the estimates based on joint estimation of $f_i^{(1)}$ and $f_i^{(2)}$ will be more efficient for the same reason that generalized least squares estimates are more efficient than ordinary least squares estimates. The second source of efficiency gain arises if there are any constraints involving both structural coefficients and covariance parameters, and if it is possible to impose these restrictions when estimating (4.10) and (4.20) jointly.

4.3.4. Further subdivision of estimation of covariance parameters

Two unattractive features of this joint estimation approach are the large number of equations involved in the implementation of GMM or 3SLS and the nonlinear parametric restrictions that must be imposed across equations when computing estimates. Fortunately, simpler estimation methods are available if a researcher is willing to estimate parameters in subsets.

Appendix B offers one approach for estimating all parameters appearing in specification (2.3) of the ARMA process underlying the U_{ti}'s, without the need to introduce any parameters associated with initial conditions. This approach replaces equation system (4.20) with an alternative set of equations exploiting relationships implied by system (4.13). In addition to reducing the number of parameters, this replacement simplifies imposition of the nonlinear restrictions inherent in relating autocorrelations. A shortcoming of this approach concerns its provision of insufficient information to develop a full specification of Θ without relying on ancillary assumptions; as demonstrated above, such a specification requires knowledge of the process generating ARMA-model initial conditions. If a researcher desires, however, to estimate only parameters of the ARMA process, then Appendix B offers a more manageable approach for conducting this estimation than use of Equations (4.12).

[28] This statement assumes that at least fourth moments of U_i exist. Homoscedasticity follows from the assumption that the U_i's are distributed independently of X_i and across individuals. Clearly $E\{f_i\} = 0$ at the true parameter values.

[29] This procedure corresponds to an estimation approach suggested by Chamberlain (1982). In Chamberlain's approach, nonlinear generalized least squares is applied to a larger model that includes the equations $f_i = \mathrm{St}\{G_i G_i' - E(G_i G_i')\}$, where the vector G_i contains all the unique elements of Y_i, Z_i and X_i.

One can achieve further simplifications in estimating parameters of error processes by developing procedures that use fitted values of U_i as dependent variables and that enable one to estimate finer subsets of parameters, such as just the AR or just the MA coefficients, using linear methods. The following subsections describe such procedures. These approaches provide especially useful diagnostic tools for inferring the basic characteristics of the underlying autocorrelation structure.

4.4. Direct estimation of autocovariances using residuals

Under conventional assumptions, econometric theory implies that one can replace the U_i's in an MM framework by their estimated residual counterparts and still obtain consistent estimates of other parameters. The residuals $\widehat{U}_i$ must be consistent estimates of the U_i for this property to hold. The problem is how to adjust the standard errors and test statistics to make this procedure of use in learning about that aspect of the error structure analyzed by the estimation approach.

A natural place to consider using residuals to form dependent variables is in system (4.13). This would eliminate the need to combine models (4.10) and (4.20) as suggested above, which involves jointly estimating numerous equations. From a methodological perspective, replacing the U_i by their fitted values amounts to fixing a subset of parameters at consistently estimated values and proceeding with estimation of another set appearing in structural equations. Proposition 4.1 identifies the conditions needed for this procedure to report asymptotically valid results for the nonfixed coefficients.

Inspection of specification (4.20) of $f_i^{(2)}$ reveals that it satisfies the property

$$E\left(\frac{\partial f_i^{(2)}}{\partial \beta_k}\right) = -E\left(\mathrm{St}\left\{\left(U_i' \otimes X_{(k)i}\right) + \left(X_{(k)i}' \otimes U_i\right)\right\}\right) = 0, \tag{4.21}$$

where β_k designates any element of the parameter vector β, and $X_{(k)i}$ constitutes the kth column vector of the matrix X_i. Define the new system of equations

$$f_i^{(3)}\left(\pi, \psi, \theta(\alpha)\right) = f_i^{(2)}\left(\pi, \psi, \hat{\beta}, \theta(\alpha)\right), \tag{4.22}$$

where $\hat{\beta}$ is a consistent estimator of β, and now specify f_i in model (3.7) as

$$f_i\left(\pi, \psi, \beta, \theta(\alpha)\right) = \left(f_i^{(1)\prime}, f_i^{(3)\prime}\right)'.$$

Given (4.21), it is evident that this specification of f_i satisfies the conditions of Proposition 4.1, with the parameter vector γ in this proposition interpreted as including π and ψ in both $f_i^{(1)}$ and $f_i^{(3)}$, β only in $f_i^{(1)}$, and $\theta(\alpha)$ in $f_i^{(3)}$, and with the parameter vector μ interpreted as including only β in $f_i^{(2)}$. Thus, as indicated by Proposition 4.1, NIV or 3SLS applied to the system of structural equations

$$\begin{pmatrix} y_i - Y_i\pi - Z_i\psi - X_i\beta \\ \mathrm{St}\{(y_i - Y_i\pi - Z_i\psi - X_i\hat{\beta})(y_i - Y_i\pi - Z_i\psi - X_i\hat{\beta})'\} - \theta(\alpha) \end{pmatrix} = \begin{pmatrix} U_i \\ V_i^* \end{pmatrix}, \tag{4.23}$$

with X_i used as instrumental variables, yields consistent estimates of π, ψ, β, and $\theta(\alpha)$ and asymptotically valid standard errors and test statistics. These estimates have the same asymptotic efficiency as the ones produced by the above joint estimation procedure, but they are easier to compute since β is not estimated in the second set of equations and parametric restrictions relating β in the two sets of equations are ignored.

When $\pi = 0$ and $\psi = 0$ – that is, when no lags in Y_{ti} and no predetermined or endogenous variables appear in Equation (2.9) – then residuals alone can be used to estimate covariance parameters. With $\widehat{U}_i = y_i - X_i\hat{\beta}$, $f_i^{(3)}$ in (4.22) becomes

$$f_i^{(3)} = \mathrm{St}\{\widehat{U}_i\widehat{U}_i'\} - \theta.$$

A corresponding formulation for this system of equations takes the form:

$$\mathrm{St}\{\widehat{U}_i\widehat{U}_i'\} = \theta + V_i^*, \tag{4.24}$$

which constitutes a seemingly unrelated regression model with $\widehat{U}_{ti}\widehat{U}_{(t-k)i}$ as dependent variables and with only intercepts as explanatory variables. Condition (4.21) and Proposition 4.1 imply that standard generalized least squares applied to model (4.24) produces consistent estimates of the elements of θ and asymptotically valid standard errors and test statistics. To estimate the covariance parameters α, nonlinear generalized least squares can be applied to model (4.24), with the functions $\theta(\alpha)$ substituted for θ.

Estimating subsets of the equations in (4.24) offers a simple framework for constructing estimates of the covariogram and the correlogram, which are valuable data analysis tools that can aid in choosing among competing specifications for the stochastic process generating the disturbances. Since estimation of the model discussed above requires only residuals, the model permits a researcher to ignore the specification of the relationships among measured variables once such a specification has been chosen, and to concentrate only on fitting the error process. Many simple tests are available for distinguishing among competing specifications. For example, with the imposition of linear constraints in subsets of equations in model (4.24), one can test whether autocovariances of a given order are constant over periods, and one can obtain a unique estimate of each autocovariance if the constancy hypothesis is accepted. Using these results, one can also perform tests for nonstationarity and other forms of heteroscedasticity. These preliminary data analyses are particularly useful for identifying the orders of the autoregressive and moving average lag polynomials and for determining whether it is reasonable to assume that the coefficients of these polynomials are constant over time.

4.5. Direct estimation of autoregressive parameters

This discussion describes a simple procedure for estimating coefficients of the autoregressive lag polynomial $a(L)$ without requiring the joint estimation of all the covariance parameters, as required in the previous approaches. This procedure offers a useful

framework for testing hypotheses that involve just the coefficients of $a(L)$. To simplify the exposition, suppose for the moment that the disturbances U_{ti} are directly observed and follow a mixed ARMA(1, 1) scheme; in particular, $U_{ti} = -a_t U_{(t-1)i} + \varepsilon_{ti} + m_1 \varepsilon_{(t-1)i}$.

This stochastic process implies that U_{ti} satisfies the linear structural equation

$$U_{ti} = -a_t U_{(t-1)i} + e_{ti}, \quad t = 3, \ldots, T, \tag{4.25}$$

where the error $e_{ti} = \varepsilon_{ti} + m_1 \varepsilon_{(t-1)i}$ follows a first-order moving-average process. The disturbance U_{1i} is predetermined for equation $t = 3$ in (4.25); the disturbances U_{1i} and U_{2i} are predetermined for equation $t = 4$, and so on, with $U_{1i}, \ldots, U_{(T-2)i}$ predetermined for equation $t = T$. To use all the available predetermined variables as instruments in the estimation of the a_t's, application of Proposition 4.2 considers expanding the system beyond (4.25) to include the structural equations:

$$\begin{aligned} &U_{(t-k)i} U_{ti} = -a_t U_{(t-k)i} U_{(t-1)i} + e^*_{ti}, \\ &\quad t = 3, \ldots, T, \ k = 2, \ldots, (t-1), \end{aligned} \tag{4.26}$$

where $e^*_{ti} = U_{(t-k)i} e_{ti}$ with $E\{e^*_{ti}\} = 0$. Formulating f_i in (3.7) according to (4.9) implies combining structural equations (4.25) and (4.26) to form f_{ti}. Applying 3SLS to this model with a constant as the only instrument in this procedure (i.e., with $Q_i = 1$ in (4.5)) produces consistent estimates of $a' = (-a_3, \ldots, -a_T)$ exploiting $U_{1i}, \ldots, U_{(t-2)i}$ as instrumental variables in the estimation of the tth equation.

Without data on U_i, combined estimation of equation systems (4.10) and (4.26) offers a framework for jointly estimating the structural coefficients and a. To translate this estimation approach into the notation of Section 3, split the vector of structural equations f_i in (3.7) into two subvectors so that $f_i = (f_i^{(1)\prime}, f_i^{(4)\prime})'$. As specified by (4.10), let $f_i^{(1)}$ denote the set of equations used in 3SLS estimation of the structural coefficients. Further, let $j = \frac{(t-3)(t-2)}{2} + (k-1)$, where $k = 2, \ldots, (t-1)$ and $t = 3, \ldots, T$. Then, with $y_{ti} - Y'_{(t-1)i}\pi - Z'_{ti}\psi - X'_{ti}\beta$ substituted for U_{ti} in Equations (4.26), form the vector $f_i^{(4)}$ whose jth element is given by

$$\begin{aligned} f^{(4)}_{ji}(\pi, \psi, \beta, a) = &\left[y_{(t-k)i} - Y'_{(t-k-1)i}\pi - Z'_{(t-k)i}\psi - X'_{(t-k)i}\beta\right] \\ &\times \left[\left(y_{ti} - Y'_{(t-1)i}\pi - Z'_{ti}\psi - X'_{ti}\beta\right)\right. \\ &\left. + a_t\left(y_{(t-1)i} - Y'_{(t-2)i}\pi - Z'_{(t-1)i}\psi - X'_{(t-1)i}\beta\right)\right]. \end{aligned} \tag{4.27}$$

Stacking (4.10) and (4.27) to obtain $f_i(\pi, \psi, \beta, a) = (f_i^{(1)\prime}, f_i^{(4)\prime})'$ creates a model in the form of (3.7). All the variables in X_i can serve as instruments for the equations incorporated in f_i since U_i is assumed to be independent of these exogenous variables. Consequently, 3SLS applied to model (3.7) with this specification of f_i yields consistent estimates of π, ψ, β and a, along with the appropriate asymptotic standard errors and test statistics.

As in the above case, joint estimation of these parameters can be simplified by evaluating β in $f_i^{(4)}$ at a consistent estimate $\hat{\beta}$, which avoids the need of imposing some

(or all) of the nonlinear parametric restrictions in estimation. Define this new system of structural equations as

$$f_i^{(5)}(\pi, \psi, a) = f_i^{(4)}(\pi, \psi, \hat{\beta}, a). \tag{4.28}$$

Differentiating the elements of $f_i^{(4)}$ in (4.27) with respect to β and computing expectations at the true parameter values yields the result

$$E\left(\frac{\partial f_i^{(4)}}{\partial \beta'}\right) = 0. \tag{4.29}$$

This finding implies that the stacked system of equations $f_i(\pi, \psi, \beta, a) = (f_i^{(1)\prime}, f_i^{(5)\prime})'$ satisfies the conditions of Proposition 4.1, with the parameter vector γ in this proposition incorporating the coefficients π and ψ in both $f_i^{(1)}$ and $f_i^{(5)}$, β only in $f_i^{(1)}$, and a in $f_i^{(5)}$, and with the parameter vector μ interpreted as including only β in $f_i^{(4)}$. Thus, according to Proposition 4.1, NIV or 3SLS applied to this specification of f_i, with all the elements of X_i used as instrumental variables, yields consistent estimates of π, ψ, β, and a and asymptotically valid standard errors and test statistics.

For those models in which $\pi = 0$ and $\psi = 0$, residuals alone can be used to estimate the autoregressive parameters a. With these parametric restrictions, one can estimate the coefficients a using only the equations in $f_i^{(5)}$. Defining $U_{ti} = y_{ti} - X_{ti}'\hat{\beta}$, these equations are

$$\widehat{U}_{(t-k)i}\widehat{U}_{ti} = -a_t\widehat{U}_{(t-k)i}\widehat{U}_{(t-1)i} + e_{ti}^*, \quad t = 3, \ldots, T,\ k = 2, \ldots, t-1. \tag{4.30}$$

Condition (4.29) and Proposition 4.1 imply that 3SLS applied to this system of equations with a constant as the only instrument produces consistent estimates of the a coefficients and asymptotically appropriate standard errors and test statistics.

Model (4.30) offers a valuable data analysis tool that can aid in determining the form of the ARMA process generating disturbances. Testing the linear constraint that $a_t = a_0$ for all t provides a simple test for the constancy of autoregressive coefficients over time. Of course, many such tests can be carried out using only subsets of the equations included in model (4.30). It is straightforward to modify this model to admit a second or higher-order autoregressive process, which provides the basis for testing for the presence of higher-order schemes. Changing the order of the moving average component of the error process alters which of the past $U_{(t-k)i}$'s are predetermined, and hence which can serve as instrumental variables for each equation. Thus, increasing the order of the moving-average process implies a reduction in the number of equations that can be included in model (4.30). Furthermore, increasing this order precludes the possibility of using this data analysis framework to estimate some period-specific values of a for the early periods of the sample.

4.6. Estimation of the partial correlation coefficients

Another useful data analysis tool found in the time series literature is the partial correlation function. The kth order partial correlation coefficient associated with the stochastic

process generating U_{ti} corresponds to the coefficient a_{kkt} in the regression equation

$$U_{ti} = -a_{k1t}U_{(t-1)i} - \cdots - a_{kkt}U_{(t-k)i} + \eta_{ti}, \tag{4.31}$$

where the error η_{ti} is uncorrelated with the regressors $U_{(t-1)i}, \ldots, U_{(t-k)i}$.

The procedures outlined in Section 4.5 provide a method for estimating the coefficients a_{kkt} for different orders and time periods. In the application of these procedures, the disturbances $U_{(t-1)i}, \ldots, U_{(t-k)i}$ are considered to be predetermined for the equation corresponding to period t, but are not necessarily predetermined for any other equation. Thus, to estimate the first-order partial correlation coefficients for periods $t = 2, \ldots, T$ using the above procedures, one would set $k = 1$ in Equation (4.26) with a_{11t} substituted for a_t.

When the U_{ti}'s represent disturbances from a regression equation (i.e., when $\pi = 0$ and $\psi = 0$ in Equation (2.9)), a more elementary approach exists for estimating the coefficient a_{kjt}. With $\widehat{U}_{ti} = y_{ti} - X'_{ti}\hat{\beta}$ denoting the least squares residuals, consider the seemingly unrelated regression model

$$\widehat{U}_{ti} = -a_{k1t}\widehat{U}_{(t-1)i} - \cdots - a_{kkt}\widehat{U}_{(t-k)i} + \eta^*_{ti}, \quad t = k+1, \ldots, T. \tag{4.32}$$

Generalized least squares applied to this model can be shown to produce consistent estimates of the coefficients a_{kjt} and appropriate large sample standard errors and test statistics.[30] Estimating these coefficients imposing equality constraints across the equations of this model (i.e., $a_{kjt} = a_{kj}$ for $t = k+1, \ldots, T$) generates a unique estimate of the kth order partial correlation coefficient and an asymptotic standard error for this coefficient. Graphing these constrained estimates of the a_{kk}'s for each value of k creates the sample partial correlation function, which is useful in the identification of time series processes.[31]

4.7. Direct estimation of moving-average parameters

This last procedure provides a method for estimating parameters associated with the moving average component of the ARMA process. These parameters include the coefficients of the lag polynomial $m(L)$ and the variance of the white noise errors, which are grouped into the parameter vector λ. There are no new concepts encountered in this estimation method.

Continuing to assume, for expositional simplicity, that an ARMA(1, 1) process generates U_{ti}, the errors $e_{ti} = \varepsilon_{ti} + m_1\varepsilon_{(t-1)i}$ appearing in Equation (4.25) capture the

[30] To verify this claim, write the seemingly unrelated regression model obtained by combining Equations (4.32) as $\omega_i = H_i\rho + \eta_i$, where $\omega_i = (\widehat{U}_{Ti}, \ldots, \widehat{U}_{(k+1)i})'$. With $\Omega = E\{\eta_i\eta'_i\}$, generalized least squares applied to this model yields the estimate $\hat{\rho} = [\sum_{i=1}^N H'_i\widehat{\Omega}^{-1}H_i]^{-1}[\sum_{i=1}^N H'_i\widehat{\Omega}^{-1}\omega_i]$. In terms of the framework of Section 3, this GLS estimate is obtained when $f_i = H_i\widehat{\Omega}^{-1}(\omega_i - H_i\rho)$ and $Q_i = 1$ in (4.5). For this specification of f_i, the conditions of Proposition 4.1 hold with $\gamma = \rho$ and with $\hat{\mu}$ interpreted to include $\hat{\beta}$ and the elements of $\widehat{\Omega}$.

[31] Section 7 presents an application illustrating this claim.

information about the moving average portion of this process. Define the vector of errors associated with Equations (4.25) as $e_i = (e_{Ti}, \ldots, e_{3i})'$. The moving-average parameters λ determine the parameterization of the variance–covariance matrix $R = E\{e_i e_i'\}$. Given data on e_i, one could estimate the elements of $r \equiv \mathrm{St}\{R\}$ by applying generalized least squares to the model $\mathrm{St}\{e_i e_i'\} = r + v_i$ – analogous to the regression model for $\mathrm{St}\{U_i U_i'\}$ given by (4.12). With the functions $r(\lambda)$ substituted for r in this model, one could estimate the parameters λ by nonlinear generalized least squares.

As in the previous analysis, jointly estimating parameters provides a method for estimating r or λ without data on e_i. Observe that

$$\begin{aligned} e_{ti} = {} & \big(y_{ti} - Y_{(t-1)i}'\pi - Z_{ti}'\psi - X_{ti}'\beta\big) \\ & + a_t\big(y_{(t-1)i} - Y_{(t-2)i}'\pi - Z_{(t-1)i}'\psi - X_{(t-1)i}'\beta\big). \end{aligned}$$

Corresponding to the system of equations included in $\mathrm{St}\{U_i U_i'\}$, form the vector of equations $f_i^{(6)}$ whose jth element is:

$$\begin{aligned} & f_{ji}^{(6)}\big(\pi, \psi, \beta, a, r(\lambda)\big) \\ & \quad = \big[\big(y_{ti} - Y_{(t-1)i}'\pi - Z_{ti}'\psi - X_{ti}'\beta\big) \\ & \qquad + a_t\big(y_{(t-1)i} - Y_{(t-2)i}'\pi - Z_{(t-1)i}'\psi - X_{(t-1)i}'\beta\big)\big] \\ & \qquad \times \big[\big(y_{(t-k)i} - Y_{(t-k-1)i}'\pi - Z_{(t-k)i}'\psi - X_{(t-k)i}'\beta\big) \\ & \qquad + a_{t-k}\big(y_{(t-k-1)i} - Y_{(t-k-2)i}'\pi - Z_{(t-k-1)i}'\psi - X_{(t-k-1)i}'\beta\big)\big], \end{aligned} \tag{4.33}$$

with $j = \frac{(T-t)(T+t-3)}{2} + (t-k-2)$ for $t = T, \ldots, 3$, and $k = (t-3), \ldots, 0$. Combining these equations along with those in $f_i^{(1)}$ and $f_i^{(4)}$ creates a model in the form of (3.7) with

$$f_i\big(\pi, \psi, \beta, a, r(\lambda)\big) = \big(f_i^{(1)\prime}, f_i^{(4)\prime}, f_i^{(6)\prime}\big)'.$$

Applying 3SLS to this model, with the exogenous variables X_i used as instruments, produces consistent estimates for π, ψ, β, a and r (or λ) and asymptotically valid standard errors and test statistics.

To reduce computational burden, this estimation procedure can be carried out with β in $f_i^{(4)}$ and $f_i^{(6)}$ replaced by a consistent estimate $\hat{\beta}$. With

$$f_i^{(7)}\big(\pi, \psi, a, r(\lambda)\big) = f_i^{(6)}\big(\pi, \psi, \hat{\beta}, a, r(\lambda)\big), \tag{4.34}$$

this approach involves 3SLS applied to model (3.7) with $f_i(\pi, \psi, \beta, a, r(\lambda)) = (f_i^{(1)\prime}, f_i^{(5)\prime}, f_i^{(7)\prime})'$ and with the elements of X_i used as instrumental variables. Condition (4.29), along with the finding

$$E\left(\frac{\partial f_i^{(6)}}{\partial \beta'}\right) = 0, \tag{4.35}$$

implies that this specification of f_i satisfies the requirements of Proposition 4.1. Consequently, application of 3SLS to this model not only yields consistent parameter estimates, but also the appropriate large sample standard errors and test statistics.

Once again, as in the previous approaches, if $\pi = 0$ and $\psi = 0$, then the system of equations in $f_i^{(1)}$ can be eliminated from the model and only data on residuals are needed to carry out estimation. With $\widehat{U}_{ti} = y_{ti} - X_{ti}'\hat{\beta}$, the vector $f_i(a, r(\lambda)) = (f_i^{(5)\prime}, f_i^{(7)\prime})'$ includes equations of the form

$$\begin{aligned}
&\widehat{U}_{(t-k)i}\widehat{U}_{ti} = -a_t\widehat{U}_{(t-k)i}\widehat{U}_{(t-1)i} + e_{ti}^*, \quad k = 2, \ldots, t-1,\ t = 3, \ldots, T,\\
&[\widehat{U}_{ti} + a_t\widehat{U}_{(t-1)i}][\widehat{U}_{(t-j)i} + a_{t-j}\widehat{U}_{(t-j-1)i}] = r_{jt} + v_{jti}^*,\\
&\quad j = (t-3), \ldots, 0,\ t = T, \ldots, 3.
\end{aligned} \tag{4.36}$$

Applying 3SLS to this system of equations using a constant as the only instrumental variable produces consistent estimates of a and r (or λ) and asymptotically valid standard errors and test statistics.

Similar to model (4.24), model (4.36) offers a relatively simple framework for constructing estimates of the covariogram and the correlogram associated with the moving average component of the error process, both of which are useful for preliminary data analysis. After one has settled on the specification of the autoregressive component, model (4.36) can potentially be useful for testing for various features of the moving-average process, such as whether it is stationary or the length of its order.

5. Estimating dynamic quantile specifications

An attractive alternative to estimating moment relationships characterizing the micro intertemporal properties of variables involves using conditional quantile regressions to describe these dynamics. This section presents a flexible empirical approach based on nonlinear instrumental variable specifications for estimating autoregressive quantile equations, exploiting the procedures outlined in the previous discussion.

5.1. Using nonlinear instrumental variable procedures to estimate quantile regressions

A familiar empirical formulation for modeling the growth of wages experienced by individuals in longitudinal data takes the form:

$$\begin{aligned}
y_{ti} &= \rho_1 y_{(t-1)i} + \cdots + \rho_r y_{(t-r)i} + X_{ti}'\beta_t + v_{ti}\\
&\equiv Y_{(t-1)i}'\rho + X_{ti}'\beta_t + v_{ti}, \quad t = 1, \ldots, T,\ i = 1, \ldots, N,
\end{aligned} \tag{5.1}$$

where y_{ti} is the dependent variable for the ith individual in the tth year, X_{ti} is a vector of exogenous measured variables, and the coefficients ρ_j and β_t are parameters. (The $t = 1$ period in (5.1) corresponds to the first period in which a researcher has data on all $y_{ti}, \ldots, y_{(t-r)i}$.) The elements of X_{ti} include exogenous variables such as year and

age effects, measures of educational attainment, and gender and race indicators. The following analysis assumes the error v_{ti} is distributed independently both over time and across individuals. Thus, the autoregressive coefficients ρ_j characterize the dynamic properties of the dependent variable after removing trends. For notational simplicity, the subsequent discussion typically ignores the i subscript on variables.

One can associate relation (5.1) with an autoregressive formulation of the κth percent quantile of the Markov distribution of y_t by imposing the restriction:

$$q_\kappa(v_{ti}|Y_{(t-1)i}, X_{ti}) = 0, \tag{5.2}$$

where $q_\kappa(\,)$ designates the κth percent quantile of the distribution of v_{ti} conditional on $Y_{(t-1)i}$ – a shorthand notation signifying all the past wages $y_{(t-1)i} \ldots y_{(t-r)i}$ appearing in (5.1) – and X_{ti}, where $\kappa \in (0, 100)$. When $\kappa = 50$, Equation (5.2) determines how the conditional median of y_t evolves over time. Although LAD procedures provide consistent estimates of the autoregressive coefficients appearing in (5.2), they have not been extensively employed.

5.1.1. Representing dynamic quantile regressions as nonlinear simultaneous equations

A class of estimators based on simultaneous equation models provides a flexible and noncumbersome procedure for estimating parameters of the dynamic quantile wage growth equation introduced above.[32] Conditioning on exogenous and predetermined instruments, this method specifies conditional quantiles of the structural error distribution as moment conditions capable of being estimated within a conventional nonlinear instrumental variables framework of the sort described in Section 3. This apparatus matches the sample analog of the conditional quantiles against their population values, employing a smoothing procedure familiar in various problems in nonparametric inference and simulation estimation. The analysis applies standard arguments to demonstrate consistency and asymptotic normality of the resulting Smoothed MM Quantile estimator.

To apply this MM quantile procedure, let y_t denote the dependent variable in year t, and let X_t denote the exogenous variables such as demographic characteristics. We are interested in obtaining information about the distribution of y_t conditional on X_t and Y_{t-1} (past values of the dependent variable). Let $q_\kappa(Y_{t-1}, X_t)$ represent the κth percent quantile of this conditional distribution, where $\kappa \in (0, 100)$. This defines the equation

$$\Pr\big(y_t < q_\kappa(Y_{t-1}, X_t)|Y_{t-1}, X_t\big) = \kappa, \tag{5.3}$$

which underlies the construction of most quantile estimation procedures. The Smoothed MM Quantile estimator uses relation (5.3) to form moment conditions. This relation implies the condition

$$E\big[1\big(y_t < q_\kappa(Y_{t-1}, X_t)\big) - \kappa|Y_{t-1}, X_t\big] = 0, \tag{5.4}$$

[32] The material presented in Sections 5.1.1 and 5.1.2 draws heavily on MaCurdy and Hong (1998).

where 1( ) represents the indicator function which takes value 1 when the condition expressed in the parentheses is true, and 0 otherwise. The indicator function inside the moment condition is neither continuous nor differentiable.

To develop a variant of this relationship capable of being specified as a moment condition in the standard framework of nonlinear method of moments estimation, observe that a smooth representation of this condition takes the form:

$$E\left[\lim_{N\to\infty} \Phi\left(\frac{y_t - q_\kappa(Y_{t-1}, X_t)}{s_N}\right) - (1-\kappa)\right] = 0, \tag{5.5}$$

where N represents the sample size, Φ is a continuously differentiable cumulative distribution function with bounded symmetric density function ϕ, and s_N is the "bandwidth" function of N that converges to 0 as N goes to ∞ at a rate slower than $N^{-1/2}$. The following analysis selects Φ to be the standard normal cumulative distribution function; a natural alternative choice would be the logit or any other cdf.

The specification of the conditional quantile function adopted in our characterization of wage dynamics is the linear distributed lag relation:

$$q_{\kappa_j}(Y_{t-1}, X_t) = \rho_1 y_{t-1} + \cdots + \rho_r y_{t-r} + X_t'\beta_t. \tag{5.6}$$

Given longitudinal data for a sample of individuals i to estimate this conditional quantile, the variant of the nonlinear simultaneous equation implied by (3.7) takes the form:

$$\begin{aligned} f_{ti} &= f_{ti}(\rho, \beta) = \Phi\left(\frac{y_{ti} - \rho_1 y_{(t-1)i} - \cdots - \rho_r y_{(t-r)i} - X_{ti}'\beta_t}{s_N}\right) - (1-\kappa_j) \\ &= U_{ti}, \end{aligned} \tag{5.7}$$

where U_{ti} is treated as a structural error with $E(U_{ti}|y_{(t-1)i}, \ldots, y_{(t-r)i}, X_{ti}) = 0$.

5.1.2. Nonlinear instrumental estimation of quantile specifications

Viewing (5.7) as a system of nonlinear simultaneous equations, application of conventional nonlinear IV or 2SLS/3SLS procedures to (5.7) yields consistent ρ and β estimates possessing large sample normal distributions. The formal proof of this proposition assumes the bandwidth parameter $s_N = N^{-d}$ for $0 < d < 1/2$.[33] One can readily verify that when $s_N \to 0$, $\Phi(\cdot)$ converges almost surely to the indicator function $1(y_t > q_\kappa(Y_{t-1}, X_t))$. Since Φ is a bounded function, one can exchange the expectation and limit operators to obtain the above smoothed moment condition. A generalized nonlinear two stage least squares estimation routine can be directly applied to this asymptotic moment condition. The estimation approach selects instrumental variables that are conditionally independent of the error terms defined by

$$1\big(y_t > q_\kappa(Y_{t-1}, X_t)\big) - (1-\kappa).$$

[33] The condition imposed on the convergence rate $0 < d < 1/2$ is needed for the proof of asymptotic normality.

The resulting Smoothed MM Quantile (SMMQ) estimators are consistent and asymptotically normally distributed with standard errors computed using robust methods. Simulation exercises reveal that this procedure accurately reproduces estimators and test statistics generated by conventional quantile estimation approaches.[34]

The selection of the value of κ_j in (5.7) determines the quantile estimated in the nonlinear IV estimation analysis. Setting $\kappa_j = \kappa_{50} = 0.5$ estimates the median, whereas setting $\kappa_j = \kappa_{25} = 0.25$ estimates the lower quartile and $\kappa_j = \kappa_{75} = 0.75$ the upper quartile. Conceptually, one can generalize specifications (5.6) and (5.7) to allow parameters to be year (or age) dependent. Estimation in this instance would require an equation for each quantile for each year (or age) that a person has current and past wage observations. Within- and cross-equation restrictions on the quantile regression coefficients could be imposed in the standard way using the multi-equation MM framework discussed below. If weighting is required to adjust for the stratified character of a dataset, then one applies the procedures summarized in Section 6.

5.2. *Jointly estimating combinations of quantile regressions*

This estimation framework extends readily to consideration of a set of quantile relations. This set may describe how a particular percentile of a distribution evolves over time, or it may summarize the relationship among several different percentiles of a conditional distribution, either in a single period or over time.

5.2.1. *Nonlinear instrumental variable estimation of quantiles in panel data*

Understanding how the jth quantile of wage rates shifts over time in a longitudinal setting involves estimating variants of (5.7) for each period available in the dataset. Allowing for the coefficients of this conditional quantile to vary over periods, the nonlinear structural equation representation of the κ_jth percentile for individual i in period t takes the form

$$g_{ti} = \Phi\left(\frac{y_{ti} - \rho_{t1}y_{(t-1)i} - \cdots - \rho_{tr}y_{(t-r)i} - X'_{ti}\beta_t}{s_N}\right) - (1-\kappa_j) = U_{ti} \qquad (5.8)$$

for $t = 1, \ldots, T$. Conditional on past wages, $y_{(t-1)i}, \ldots, y_{(t-r)i}$, and X_{ti}, the error U_{ti} has mean 0. Constraining the coefficients $\rho_{tm} = \rho_m$ for all t yields a parameterization analogous to a conventional ARMA$(r, 0)$ process.

With the structural errors U_{ti} distributed independently over time as well as across individuals, nonlinear 2SLS offers a simple procedure for estimating the coefficients ρ_j and β – assumed here to be constant over time. This approach treats each g_{ti} in (5.8) as a separate observation, with the sample made up of all the combinations of (t, i)

[34] MaCurdy and Hong (1998) explore the performance of various choices for the bandwidth parameter in a simulation study; the estimation analysis below relies on the results of this exercise.

where $t = 1, \ldots, T$ and $i = 1, \ldots, N$. The instrumental variables used in estimation consist of functions of $y_{(t-1)i}, \ldots, y_{(t-r)i}$ and X_{ti}. Expressed in terms of the notation of Section 3, this estimation procedure amounts to setting

$$f_i = g_{tj} \quad \text{and} \quad Q_i = \{\text{functions of } y_{(t-1)j}, \ldots, y_{(t-r)j} \text{ and } X_{tj}\} \text{ where } t = 1, \ldots, T \text{ and } j = 1, \ldots, N. \tag{5.9}$$

This formulation for f_i substitutes for (3.7), with the index i merely counting all observations obtained by stacking the time series observations for all individuals. If one suspects the errors U_{ti} are heteroscedastic, then nonlinear 2SLS estimation should be implemented calculating robust standard errors corresponding to the asymptotic distribution (3.11).

Alternatively, if one wishes to allow for an individual's U_{ti} errors to be dependent in some way over time, with U_{ti}'s still being independent across individuals, then multiequation methods incorporating predetermined variables described in Sections 3.2.1 and 4.1.3 offer an approach for estimating the parameters of (5.8). The predetermined variables include the past values of wages, so the analog of regression equation (4.7) becomes

$$y_{ti} = \delta_t Q_i + \eta_{ti}, \quad t = 1, \ldots, T. \tag{5.10}$$

The regressors Q_i incorporate all of the exogenous variables of the model, including the relevant functions of the components making up X_i. The implied formulation for f_{ti} in this case is

$$f_{ti}(\gamma) = \begin{pmatrix} g_{ti} \\ \hat{\eta}_{(t-1)i} g_{ti} \\ \vdots \\ \hat{\eta}_{(t-r)i} g_{ti} \end{pmatrix}, \quad t = 1, \ldots, T. \tag{5.11}$$

The parameters γ include all the coefficients $\rho_{t1}, \ldots, \rho_{tr}, \beta_t$ for $t = (r+1), \ldots, T$. (These coefficients may be constrained.) Relation (4.9) gives the specification for f_{ti}. In constructing f_{ti}, one can replace the $\hat{\eta}_{ti}$ with functions of the $\hat{\eta}_{ti}$. Moreover, one can expand f_{ti} to include additional elements involving extra functions of the $\hat{\eta}_{ti}$. Formulas (3.21) and (3.22) give the optimal specification for the instrumental variables.

5.2.2. *Estimating dynamic specifications describing several quantiles*

Suppose a researcher wishes to estimate more than one quantile to describe the evolution of the wage distribution more fully. In particular, suppose interest focuses on estimating the J quantiles $0 < \kappa_1 < \kappa_2 < \cdots < \kappa_J < 1$. The system of J nonlinear simultaneous equations providing for the estimation of these percentiles takes the form

$$g_{jti} = \Phi\left(\frac{y_t - q_{\kappa_j t}(Y_{t-1}, X_t)}{s_N}\right) - (1 - \kappa_j) = 0, \quad j = 1, \ldots, J. \tag{5.12}$$

These equations apply to period t. Introducing a variant of system (5.12) for each available period in panel data permits an analysis of how these J quantiles shift over time.

Each of these equations can be separately estimated using single equation two-stage least square methods. To improve efficiency given the available instruments, one can apply a three-stage nonlinear least squares or joint-equation MM estimation procedure by weighting the J equations optimally. Under the conditions noted briefly in Section 5.1.2,[35] the nonlinear instrumental variable procedures presented in Sections 3 and 4 applied to (5.12) produce consistent estimates and valid asymptotic distributions for the coefficients of the quantile functions. The optimal weighting matrix is determined by the variance–covariance matrix of J *sign-variables* defined by

$$1\big[y_t > q_{\kappa_j t}(Y_{t-1}, X_t)\big], \quad j = 1, \ldots, J. \tag{5.13}$$

This matrix depends only on the κ's associated with the specific distribution of the error term. In particular, $\mathrm{Var}[1(y_t > q_\kappa(Y_{t-1}, X_t)] = \kappa(1-\kappa)$, and for $\kappa_p > \kappa_j$,

$$\begin{aligned}&\mathrm{Cov}\big[1\big(y_t > q_{\kappa_p}(Y_{t-1}, X_t)\big), 1\big(y_t > q_{\kappa_j}(Y_{t-1}, X_t)\big)\big]\\ &\quad = 1 - \kappa_p - (1-\kappa_p)(1-\kappa_j).\end{aligned} \tag{5.14}$$

One can permit flexible and unknown forms of heteroscedasticity in calculating the optimal weighting matrix used in MM estimation. Incorporating these generalizations involves implementing the conventional approach utilized in multiple-equation MM procedures.

6. Use of sample weights and unbalanced data

When and how to weight data are two of the most important and least understood concepts in estimation. The subsequent discussion begins with the basic principles underlying weighting, and then summarizes how these basics apply to MM estimation with longitudinal data with nonlinear specifications. The discussion documents how one must modify MM formula to account for stratified sampling. The section ends by describing a modified weighting-type procedure enabling researchers to use conventional methods to estimate intertemporal specifications using unbalanced datasets – datasets not supplying a perfect overlap in the time periods for individuals included in the cross-sectional samples.

6.1. Basics of weighting to account for stratified sampling

Practically all micro data are collected using a stratified sampling frame, meaning that observations with particular characteristics are drawn in proportions differing from

[35] See MaCurdy and Hong (1998) for more details.

those of the true population. Throughout this section, the discussion considers households as observations, but the weighting procedures outlined here obviously apply for whatever observation unit happens to be relevant for an analysis, such as individuals or firms. The true population refers to the group whose distribution a researcher wishes to discern.

Suppose one would like to infer the mean of a variable y, say income, in a population with households of two types: Type 1 and Type 2. Type 1 may refer to a poor household, and Type 2 to a nonpoor household; alternatively, Type 1 may designate a black family, whereas Type 2 indicates a white one. In the true population, assume Type 1 households make up proportion P_1 of the population, and Type 2 households constitute the remaining $P_2 = (1 - P_1)$ proportion. Thus, P_1 represents the probability that a randomly drawn household from the true population is Type 1. With y_i denoting the value of y for household i, suppose the expected value of y differs for the two types of household with

$$E(y_i|\text{Type 1}) = \mu_1, \qquad E(y_i|\text{Type 2}) = \mu_2.$$

Thus, the mean of y in the population is

$$E(y_i) \equiv \mu = \mu_1 P_1 + \mu_2 P_2. \tag{6.1}$$

A stratified sample includes observations on household types in proportions that differ from P_1 and P_2. Data collectors may want an oversample of poor or black families to enable them to learn about the circumstances of these groups with added precision. Suppose this oversample occurs for Type 1 households; out of a sample of size N, N_1 are of Type 1 with the sample share $S_1 \equiv N_1/N > P_1$. The sample average of y equals

$$\begin{aligned}\bar{y} = \frac{1}{N}\sum_{1}^{N} y_i &= S_1 \frac{1}{N_1} \sum_{\{i \in \text{Type } 1\}} y_i + S_2 \frac{1}{N_2} \sum_{\{i \in \text{Type } 2\}} y_i \\ &= S_1 \hat{\mu}_1 + S_2 \hat{\mu}_2, \end{aligned} \tag{6.2}$$

where $S_2 \equiv (N - N_1)/N \equiv N_2/N$, and

$$\hat{\mu}_j = \frac{1}{N_j} \sum_{\{i \in \text{Type } j\}} y_i, \quad j = 1, 2.$$

The sample mean $\hat{\mu}_j$ calculated over Type j households in the sample consistently estimates the expected value μ_j. Since the sample shares S_1 and S_2 do not converge asymptotically to the true population shares P_1 and P_2, $\bar{y}$ clearly does not consistently estimate μ.

Weighting the data solves this problem. Define the weight for observation i as

$$w_i = \frac{P_j}{S_j}, \tag{6.3}$$

where j signifies household i's type. Weighting observations and recalculating the sample mean yields

$$\begin{aligned}\bar{y}_w &= \frac{1}{N}\sum_{1}^{N} w_i y_i = S_1 w_1 \frac{1}{N_1}\sum_{\{i\in\text{Type }1\}} y_i + S_2 w_2 \frac{1}{N_2}\sum_{\{i\in\text{Type }2\}} y_i \\ &= P_1\hat{\mu}_1 + P_2\hat{\mu}_2 = \hat{\mu}. \end{aligned} \tag{6.4}$$

The use of a weight inflates observations that are under-represented and deflates values associated with over-represented households.

6.2. Weighting to account for more sophisticated sample stratification

One can readily generalize the above scheme to admit many types of households. Suppose the vector of characteristics X_i designates a household i's type, and the share of this type in the overall population equals $P_i = P(X_i)$. The expected value of y in this population is

$$\mu = \sum_{i=1}^{N} E(y_i|X_i)P(X_i) = \sum_{i=1}^{N} \mu_i P_i.$$

With S_i representing the share of Type X_i in the sample, the weight for observation i with these characteristics is

$$w_i = \frac{P_i}{S_i}. \tag{6.5}$$

Computing the average of the sample using weights (6.5) yields

$$\begin{aligned}\bar{y}_w = \frac{1}{N}\sum_{i=1}^{N} w_i y_i &= \sum_j S_j w_j \left[\frac{1}{N_j}\sum_{\{i\in\text{Type }j\}} y_i\right] \\ &= \sum_j P_j\hat{\mu}_j \\ &= \sum_{i=1}^{N} P_i y_i = \hat{\mu}. \end{aligned} \tag{6.6}$$

This relation generalizes (6.4). In the first and second lines of (6.6), the index j designates types. The last line presumes each observation is its own type, producing the most general form of weighting.

6.2.1. Typical form of weights provided in survey data

Datasets often report weights in a way such that adding weights for observations of a particular type X_i yields the total number of households of that type in the overall

population. Thus, in place of (6.5), datasets may provide the weight:

$$\begin{aligned} w_i^* &= \frac{1}{N_i} P_i \cdot (\text{Total number of households in true population}) \\ &= \frac{1}{N} w_i \cdot (\text{Total number of households in true population}). \end{aligned} \tag{6.7}$$

The i subscript in these expressions designates observation i's type. Summing these weights over all members of types in the set K yields

$$\begin{aligned} \sum_{\{i \in K\}} w_i^* &= \sum_{\{j \in K\}} \frac{1}{N} w_j \left[\sum_{\{i \in \text{Type } j\}} 1 \right] \cdot (\text{Total number in true population}) \\ &= \sum_{\{j \in K\}} \frac{1}{N} w_j N_j \cdot (\text{Total number in true population}) \\ &= \sum_{\{j \in K\}} P_j \cdot (\text{Total number of households in true population}) \\ &= \text{Total number of households in true population of types included} \\ &\quad \text{in the set } K. \end{aligned} \tag{6.8}$$

Thus, when the set K includes all types j, this quantity measures the total number of households in the true population; equivalently,

$$\sum_{\{i \in K\}} w_i^* = \sum_{i=1}^{N} w_i^* = \text{Total number of households in true population.}$$

Computing weighted averages of the observations y_i to estimate expected values of y in the true population takes the form

$$\frac{1}{N} \sum_{\{i \in K\}} w_i y_i = \frac{\sum_{\{i \in K\}} w_i^* y_i}{\sum_{\{i \in K\}} w_i^*} = \sum_{\{i \in K\}} P_i y_i = \hat{\mu}. \tag{6.9}$$

The last step in (6.9) treats every individual as his/her own type, with K covering all possible types. The standard error of this estimated mean is the square root of the quantity

$$\begin{aligned} \frac{1}{N^2} \sum_{\{i \in K\}} [w_i y_i - \hat{\mu}]^2 &= \frac{1}{N^2} \sum_{\{i \in K\}} w_i^2 y_i^2 - \frac{1}{N} \hat{\mu}^2 \\ &= \frac{\sum_{\{i \in K\}} [w_i^* y_i]^2}{[\sum_{\{i \in K\}} w_i^*]^2} - \frac{1}{N} \left[\frac{\sum_{\{i \in K\}} w_i^* y_i}{\sum_{\{i \in K\}} w_i^*} \right]^2. \end{aligned} \tag{6.10}$$

Most software packages use these formulas to compute weighted means and their associated standard errors.[36]

[36] Software packages often present several options for using different forms of weights.

6.2.2. Calculating statistics for subpopulations

Instead of representing the entire population, suppose the set K includes only a subset of household types. Let μ_K designate the mean of the subpopulation comprised of all household types making up the set K. A consistent estimate of this mean takes the form

$$\begin{aligned}\frac{\sum_{\{i\in K\}} w_i^* y_i}{\sum_{\{i\in K\}} w_i^*} &= \frac{\sum_{i=1}^N w_i^*}{\sum_{\{i\in K\}} w_i^*} \sum_{\{i\in K\}} \frac{1}{N} w_i y_i \\ &= \frac{\sum_{i=1}^N w_i^*}{\sum_{\{i\in K\}} w_i^*} \sum_{\{j\in K\}} S_j w_j \left[\frac{1}{N_j} \sum_{\{i\in \text{Type } j\}} y_i\right] \\ &= \frac{\sum_{i=1}^N w_i^*}{\sum_{\{i\in K\}} w_i^*} \sum_{\{j\in K\}} P_j \hat{\mu}_j \\ &= \hat{\mu}_K. \end{aligned} \tag{6.11}$$

In this weighted average, note that the expression

$$\frac{\sum_{i=1}^N w_i^*}{\sum_{\{i\in K\}} w_i^*} P_j$$

corresponds to the proportion of Type j households that make up the true subpopulation defined by set K.

The variable y_i in the above discussion can represent any general quantity of the data, including higher-order terms, allowing the estimation of higher order moments of y. This analysis merely shows how to estimate moments associated with the true population using observations from a stratified sample. As one example, a consistent estimate of the variance of y_i in subpopulation K is:

$$\widehat{E}\big(y^2|\{i\in K\}\big) - \big[\widehat{E}\big(y|\{i\in K\}\big)\big]^2 = \frac{\sum_{\{i\in K\}} w_i^* [y_i]^2}{\sum_{\{i\in K\}} w_i^*} - \left[\frac{\sum_{\{i\in K\}} w_i^* y_i}{\sum_{\{i\in K\}} w_i^*}\right]^2.$$

Note, in contrast to the classic unweighted case – which corresponds to the situation $w_i^* = 1$ for all i – the square of the standard error of the weighted estimated mean is not proportional to the estimated variance of y_i. Weighting alters the variability of the sample average.

6.3. Weighting in method-of-moments procedures to compute estimators

Nonlinear functions of variable and parameter vectors can also be represented by the y_i variable in the above discussion. The analysis merely shows how to estimate moments associated with the true population using observations from a stratified sample. An important question concerns how these lessons can be implemented using the general estimation procedures described in Section 3.

As noted in Section 3, MM procedures minimize a distance function of the form (3.2) to compute estimates ($\tilde{\gamma}$) for the parameters γ. The quantities $L_N(\gamma)$, $V_N(\gamma)$, and $S_N(\gamma)$ appearing in (3.2) and (3.3), which specify the asymptotic distribution of $\tilde{\gamma}$, take the form

$$L_N(\gamma) \equiv \frac{1}{N}\sum_{i=1}^{N} \ell_i(\gamma), \quad S_N \equiv \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell_i}{\partial \gamma'} \quad \text{and}$$

$$V_N(\gamma) \equiv \frac{1}{N}\sum_{i=1}^{N} \ell_i(\gamma)\ell_i(\gamma)'.$$

As in the previous analysis, the matrix H_N appearing in distance function (3.2) and the asymptotic distribution (3.3), is any positive definite matrix.

The asymptotic properties of the estimator $\tilde{\gamma}$ critically rely on $L_N(\gamma_0) \xrightarrow{s} 0$, and this convergence property in turn relies on the sample average of the $\ell_i(\gamma)$'s converging to an expectation of zero based on the true distribution. Thus, to have average (3.1) converge to the appropriate expectation in the case of a stratified sample, one replaces $L_N(\gamma)$ by its weighted counterpart

$$L_N(\gamma) \equiv \frac{\sum_{\{i\in K\}} w_i^* \ell_i(\gamma)}{\sum_{\{i\in K\}} w_i^*} = \frac{1}{N}\sum_{\{i\in K\}} w_i \ell_i(\gamma), \tag{6.12}$$

where the set K includes all types included in the overall dataset. The corresponding formulations for $S_N(\gamma)$ and $V_N(\gamma)$ take the form

$$S_N(\gamma) \equiv \frac{\sum_{\{i\in K\}} w_i^* \frac{\partial \ell_i}{\partial \gamma'}}{\sum_{\{i\in K\}} w_i^*}, \qquad V_N(\gamma) \equiv \frac{1}{N}\frac{\sum_{\{i\in K\}} w_i^{*2} \ell_i(\gamma)\ell_i(\gamma)'}{[\sum_{\{i\in K\}} w_i^*]^2}. \tag{6.13}$$

With (6.12) used to construct the distance function (3.2) with any positive-definite matrix H_N, the extremum estimator $\tilde{\gamma}$ consistently estimates the value of γ_0 associated with the true population. Moreover, $\tilde{\gamma}$ possesses asymptotic distribution (3.3) with $\widetilde{V}_N \equiv V_N(\tilde{\gamma})$ and $\widetilde{S}_N \equiv S_N(\tilde{\gamma})$, where (6.13) gives the formula for $V_N(\gamma)$ and $S_N(\gamma)$.

The generalized least-squares variant of the distance function providing for the computation of the most efficient method-of-moments estimator $\hat{\gamma}$ is still the function $C(\gamma)$ specified by (3.4). The quadratic form matrix in $C(\gamma)$ is the inverse of $\widetilde{V}_N \equiv V_N(\tilde{\gamma})$, with (6.13) again giving the formula for $V_N(\gamma)$, and $\tilde{\gamma}$ being any consistent weighted estimator. The extremum estimator $\hat{\gamma}$, the value of γ minimizing $C(\gamma)$, consistently estimates the value γ_0 associated with the true population. Moreover, $\hat{\gamma}$ follows an asymptotic distribution given by (3.5) with (6.13) giving the formulas for $V_N(\gamma)$ and $S_N(\gamma)$.

6.4. Weighting in LS and instrumental variable procedures to compute estimators

What are the implications of applying the above weighting procedures for implementing least squares? For nonlinear least squares? For nonlinear 3SLS?

6.4.1. Familiar form of weighting in LS procedures

For least squares, consider the simple linear model

$$y_i = X_i'\gamma + \xi_i, \quad i = 1, \ldots, N.$$

Expressed in terms of equation system (3.7), this relation translates as

$$f_i = y_i - X_i'\gamma = \xi_i. \tag{6.14}$$

Least squares amounts to selecting the instrumental variables $Q_i = X_i$, and minimizing distance function (3.2) with

$$\begin{aligned} L_N(\gamma) &\equiv \frac{1}{N}\sum_{i=1}^{N} \ell_i(\gamma), \quad \text{where } \ell_i(\gamma) = X_i\big(y_i - X_i'\gamma\big), \\ H_N &\equiv \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1}. \end{aligned} \tag{6.15}$$

The implied formulations for the matrices $\widetilde{S}_N$ and $\widetilde{V}_N$ are:

$$\widetilde{S}_N = -\left[\frac{1}{N}\Sigma_i\big(X_i X_i'\big)\right], \qquad \widetilde{V}_N = \left[\frac{1}{N}\Sigma_i\big(X_i X_i'\tilde{\xi}_i^2\big)\right]. \tag{6.16}$$

Since the system of equations $L_N = 0$ fully defines the least squares estimator $\hat{\gamma}$ (i.e., the number of equations equals the number of elements estimated in γ), asymptotic distribution (3.6) approximates the large-sample distribution of $\hat{\gamma}$. The formula for the variance–covariance matrix in (3.6) provides for the computation of robust standard errors. Obviously, under the assumption of homoscedasticity, this formula simplifies to the familiar least squares specification.

Weighting to account for stratified sampling amounts to computing a least squares estimator (or generalized least squares estimator) for the equation

$$\sqrt{w_i}y_i = \sqrt{w_i}X_i'\gamma + \sqrt{w_i}\xi_i.$$

Considered in the context of the general estimation approach described in Section 3.1, this estimation procedure sets

$$f_i = \sqrt{w_i}\big[y_i - X_i'\gamma\big] = \sqrt{w_i}\xi_i, \tag{6.17}$$

and selects the instrumental variables $Q_i = \sqrt{w_i}X_i$. The weighted least squares estimator $\hat{\gamma}_w$, minimizes the distance function (3.2) with

$$\begin{aligned} L_N(\gamma) &\equiv \frac{1}{N}\sum_{i=1}^{N} \ell_i(\gamma), \quad \text{where } \ell_i(\gamma) = w_i X_i\big(y_i - X_i'\gamma\big), \\ H_N &\equiv \left[\frac{1}{N}\sum_{i=1}^{N} w_i^2 X_i X_i'\right]^{-1}. \end{aligned} \tag{6.18}$$

The implied formulations for the matrices $\widetilde{S}_N$ and $\widetilde{V}_N$ are:

$$\widetilde{S}_N = -\left[\frac{1}{N}\Sigma_i\left(w_i X_i X_i'\right)\right], \qquad \widetilde{V}_N = \left[\frac{1}{N}\Sigma_i\left(w_i^2 X_i X_i' \tilde{\xi}_i^2\right)\right]. \tag{6.19}$$

The specification for L_N in (6.18) clearly possesses the form required by (6.12) to adjust for the use of a stratified sample, so $\hat{\gamma}_w$ consistently estimates the value of γ associated with the true population. Asymptotic distribution (3.6) approximates the large-sample distribution of $\hat{\gamma}_w$, with (6.19) serving as the components appearing in the formula for the variance–covariance matrix of this distribution.

Although the variance–covariance formula agrees with what one would obtain through computing robust standard errors by applying least squares to estimate the weighted regression equation (6.17), be aware that this formula is *not* the one typically reported by weighted regression software packages. These packages presume that weighting is done to induce homoscedasticity, which renders the matrix $\widetilde{V}_N$ proportional to $\widetilde{S}_N$ with the factor of proportionality consistently estimated by

$$\tilde{\sigma}_N^2 = \left[\frac{1}{N}\Sigma_i\left(w_i \tilde{\xi}_i^2\right)\right].$$

This simplification relies on the assumption that $\text{Var}(\xi_i)$ is proportional to w_i. There is little reason to believe this relationship holds in a stratified sample, for the sampling weights are not designed with this consideration in mind. For example, it is possible for a stratified sample to have different means but the same variances across groups. In this case, weighting is needed to compute means for the overall population, but a simple average estimates variances. Weighting in this instance *induces* heteroscedasticity. The variance–covariance formula appearing in (3.5), with (6.19) inserted as the components of this expression, consistently estimates the appropriate standard errors and test statistics regardless of how weighting alters the variances of disturbances.

6.4.2. Weighting with LS interpreted as an IV procedure

Alternatively, representing the weighted least squares estimators within a 2SLS framework offers an approach for computing the appropriate standard errors when one is willing to assume the regression errors ξ are homoscedastic across observations. Designate the specification for f_i by (6.14), and implement 2SLS selecting the instrumental variables $Q_i = w_i X_i$. The application of 2SLS minimizes distance function (3.2) with

$$\begin{aligned} L_N(\gamma) &\equiv \frac{1}{N}\sum_{i=1}^{N} \ell_i(\gamma), \quad \text{where } \ell_i(\gamma) = w_i X_i\left(y_i - X_i'\gamma\right), \\ H_N &\equiv \left[\frac{1}{N}\sum_{i=1}^{N} w_i^2 X_i X_i'\right]^{-1}. \end{aligned} \tag{6.20}$$

Relations (6.19) give the implied formulations for the matrices $\widetilde{S}_N$ and $\widetilde{V}_N$. As mentioned before, the specification of H_N is irrelevant in the calculation of estimates with

the form of L_N given by (6.18) – in the terminology of 2SLS, this equation is exactly identified. So, the weighted least squares estimator $\hat{\gamma}_w$ is the 2SLS estimator. Asymptotic distribution (3.6) once again approximates the large-sample distribution of $\hat{\gamma}_w$, with (6.19) serving as the components appearing in the formula for the variance–covariance matrix of this distribution. Carrying out 2SLS estimation with the robust standard error option selected to calculate standard errors uses this formula for the variance–covariance matrix. If one assumes the errors ξ are homoscedastic, then the conventional standard error formula for 2SLS consistently estimates the variance–covariance matrix of (3.6).

This 2SLS representation of weighted least squares readily accommodates the nonlinear regression case. One computes the weighted nonlinear least squares estimator using nonlinear 2SLS procedures by specifying $f_i = y_i - g(X_i, \gamma)$, where g is a known nonlinear function, and selecting the instrumental variables as [37]:

$$Q_i = w_i \frac{\partial g_i}{\partial \gamma}\bigg|_{\tilde{\gamma}}.$$

With these specifications of f_i and Q_i, all of the above findings apply in computing the weighted nonlinear least squares estimator and its asymptotic distribution.

6.4.3. Weighting in nonlinear IV procedures

Now consider NIV and nonlinear 3SLS estimation with weighting, which encompass all other linear and nonlinear MM procedures. Application of weighted NIV to the model specified in system (3.7) selects the instrumental variables

$$Q_{ji} = w_i G_{ji}, \tag{6.21}$$

where G_{ji} represents the instrumental variables a researcher would use in the absence of weighting. The implied formulation for the ℓ_i's in the NIV framework takes the form

$$\begin{aligned} L_N(\gamma) &\equiv \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \ell_{Ti}(\gamma) \\ \vdots \\ \ell_{1i}(\gamma) \end{pmatrix} = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} w_i G_{Ti} f_{Ti} \\ \vdots \\ w_i G_{1i} f_{1i} \end{pmatrix} \equiv \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} Q_{Ti} f_{Ti} \\ \vdots \\ Q_{1i} f_{1i} \end{pmatrix} \\ &\equiv \frac{1}{N}\sum_{i=1}^{N} \Delta_i f_i, \end{aligned} \tag{6.22}$$

where the matrix Δ_i is given by (3.9). (Relation (6.22) presumes consideration of only a single structural equation per period for expositional simplicity; inclusion of Kronecker products in forming the ℓ_i's as in (3.8) will generalize this expression to permit consideration of multiple structural equations per period.) The specification of L_N implied by (6.22) clearly possesses the form required by (6.12) to adjust for the use of a stratified

[37] Note, this specification of Q_i corresponds to the weighted value of the optimal choice of instrumental variables presented in (3.13).

sample, so the value of $\hat{\gamma}_w$ minimizing distance function (3.10) consistently estimates the value of γ_0 associated with the true population. Moreover, the weighted estimator $\hat{\gamma}_w$ possesses asymptotic distribution (3.11). If the homoscedasticity assumption applies for the structural errors ξ_i, then $\hat{\gamma}_w$ can be interpreted as the nonlinear 3SLS estimator obtained by minimizing distance function (4.3), and (4.4) gives the asymptotic distribution of $\hat{\gamma}_w$. An equivalent representation has the 3SLS estimator $\hat{\gamma}_w$ minimize distance function (4.5) with its asymptotic distribution given by (4.6), where the quantities $w_i Q_i$ replace Q_i in these expressions.

6.5. Which weights should be used in longitudinal analyses?

The selection of weights appropriate for a panel data exercise requires the following steps: first, a decision of exactly which population a researcher wants to emulate; and, second, a clear understanding of what the weights are intended to represent. Documentation accompanying survey data on weights seldom discuss their use other than vaguely noting the broad category of the population the weights are meant to replicate. For example, there are usually weights for estimating relationships involving observations on individuals, and weights for families. The choice among these options is usually obvious since one emulates the population of individuals for some region or age range, and the other models a population of families or households. However, there are often different sets of weights for each year as well, leaving the question of which to use.

To discuss the principles underlying the answer to this question, return to the problem of estimating the expected value of a variable y_{ti}, where i refers to a household and t designates the year. The variable y may represent a simple variable such as income, or it may equal the product of income in the current period and some previous period. To estimate the mean of y in the target population using the framework outlined above, the specification f_{ti} in equation system (3.7) takes the form

$$f_{ti} = y_{ti} - \gamma = \xi_{ti}, \quad t = 1, \ldots, T, \ i = 1, \ldots, N. \tag{6.23}$$

With G_{ti} again representing the instrumental variables used ignoring weighting, replace the weighted variant of instrumental variables appearing in (6.22) by the quantity

$$Q_{ti} = w_{ti} G_{ti}, \tag{6.24}$$

with the specification of w_{ti} designated below. In the estimation of Equation (6.23), the constant $G_{ti} = 1$ constitutes the only instrumental variable. The implied formulation for the ℓ_i's in the MM framework becomes

$$\begin{aligned} L_N(\gamma) &\equiv \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} \ell_{Ti}(\gamma) \\ \vdots \\ \ell_{1i}(\gamma) \end{pmatrix} = \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} w_{Ti} G_{Ti} f_{Ti} \\ \vdots \\ w_{1i} G_{1i} f_{1i} \end{pmatrix} \equiv \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} Q_{Ti} f_{Ti} \\ \vdots \\ Q_{1i} f_{1i} \end{pmatrix} \\ &\equiv \frac{1}{N} \sum_{i=1}^{N} \Delta_i f_i, \end{aligned} \tag{6.25}$$

where relation (3.9) gives the formula for the matrix Δ_i. Implementing the weighted NIV procedure described above produces consistent estimates for the value of γ_0 associated with the true population assuming the weights correspond to this population.

So how does one select the w_{ti}? Rarely would one select the w_{ti} to be the annual weights provided in survey longitudinal data. The annual weights are meant to adjust for the fact that households attrit from the sample over time, and this attrition does not occur randomly across household types. Moreover, these weights may also adjust to recognize that, for example, the national population changes due to immigration. These adjustments recognize that what was representative in one year is not representative a decade later.

Most empirical analyses investigating intertemporal relationships work with balanced data, meaning that all observations i are deleted from the dataset if any of their observations (t, i) are missing. Suppose a researcher wishes to estimate relationships over the period 1 to T, and uses all data on households i who are in the data in period 1 and still part of the dataset through period T. Conceptually, the weight w_{Ti} would be the proper one to use in such analyses; so, in (6.24) and (6.25) one would set $w_{ti} = w_{Ti}$ for all t. This selection would also be appropriate if the researcher conducted the analysis starting after period 1.

The validity of this choice, of course, critically depends on the weights properly adjusting for attrition. The circumstances under which weights accomplish this task rely on accounting for the potential presence of endogenous sample selection, an exceedingly complicated problem that requires implementation of a structural analysis to infer corrected probabilities of sample inclusion. This significant topic is beyond the scope of the chapter.[38] Additional problems can arise if the annual weights in longitudinal data attempt to adjust for recent-arrival immigrant households who were underrepresented in the original sample. For the weights to adjust properly for stratification, a researcher must presume that households designated to be equivalent by sample weights do not differ depending on whether they are original-sample or recent-arrival immigrant households.

6.6. Estimation with unbalanced data

What about using all observations available in panel data to estimate relationships, irrespective of whether these observations come from households who were not represented in some years? Using an unbalanced dataset requires adjustments to the output of test statistics reported by conventional estimation procedures.

6.6.1. Characterizing estimators computed using unbalanced data

Suppose a data source offers observations on N households i for some years t during the period $1, \ldots, T$. In year t, observations exist on all households who are members of

[38] The creation of weights in such instances is closely linked to the theory underlying choice based sampling, a topic touched upon in the Handbook chapter by McFadden (1984).

the set $\{i \in \{t\}\}$; there are N_t such households in year t. A particular household j may be represented in a combination of years, implying it may be in any combination of the sets $\{i \in \{1\}\}, \ldots, \{i \in \{T\}\}$. A stratified sample assigns a household missing in year t a weight equal to 0; that is, $w_{ti} = 0$ when an observation on i is unavailable for year t. A convenient reformulation of the weights useful in the subsequent analysis takes the form

$$W_{ti} = \frac{N}{N_t} w_{ti}. \tag{6.26}$$

Often the weights provided in data sources are in fact W_{ti} and not w_{ti}, for they adjust for the smaller sizes of the cross-sectional samples. Generally, the weight w_{ti} appearing in (6.26) refers to the cross-sectional weight applicable for year t; this selection of w_{ti} presumes that the most recent endogenous variable included in the equation weighted by w_{ti} is from period t. A household i with a missing observation in period t has $W_{ti} = 0$. One need not distinguish between W_{ti} and w_{ti} in the formulation of the w^* weights (representing population sizes) discussed in Section 6.2.

An MM estimator using unbalanced data minimizes a distance function of the form (3.2) where

$$\begin{aligned} L_N(\gamma) &= \begin{pmatrix} \frac{1}{N_T} \sum_{i \in \{T\}} \ell^*_{Ti}(\gamma) \\ \vdots \\ \frac{1}{N_1} \sum_{i \in \{1\}} \ell^*_{1i}(\gamma) \end{pmatrix} = \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} \frac{N}{N_T} \ell^*_{Ti}(\gamma) \\ \vdots \\ \frac{N}{N_1} \ell^*_{1i}(\gamma) \end{pmatrix} \\ &\equiv \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} \ell_{Ti}(\gamma) \\ \vdots \\ \ell_{1i}(\gamma) \end{pmatrix}. \end{aligned} \tag{6.27}$$

This relation defines the vectors ℓ^*_{ti} and ℓ_{ti} so they differ only by the factor N/N_t, which enlarges values of ℓ^*_{ti} to account for summing over a greater number of households than there are observations in year t. When an observation on household i is missing in year t, $\ell_{ti} = 0$ as is the case for the weight W_{ti}. In the case of NIV estimation with stratified samples, (6.27) becomes

$$\begin{aligned} L_N(\gamma) &= \begin{pmatrix} \frac{1}{N_T} \sum_{i \in \{T\}} \ell^*_{Ti}(\gamma) \\ \vdots \\ \frac{1}{N_1} \sum_{i \in \{1\}} \ell^*_{1i}(\gamma) \end{pmatrix} = \begin{pmatrix} \frac{1}{N_T} \sum_{i \in \{T\}} w_{Ti} G_{Ti} f_{Ti} \\ \vdots \\ \frac{1}{N_1} \sum_{i \in \{1\}} w_{1i} G_{1i} f_{1i} \end{pmatrix} \\ &= \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} \frac{N}{N_T} w_{Ti} G_{Ti} f_{Ti} \\ \vdots \\ \frac{N}{N_1} w_{1i} G_{1i} f_{1i} \end{pmatrix} \equiv \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} W_{Ti} G_{Ti} f_{Ti} \\ \vdots \\ W_{1i} G_{1i} f_{1i} \end{pmatrix}, \end{aligned} \tag{6.28}$$

where G_{ji} again constitutes the instrumental variables a researcher would use in the absence of weighting. Expressed in terms of the notation of Section 3, (6.28) trans-

lates to

$$L_N(\gamma) = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} W_{Ti}G_{Ti}f_{Ti} \\ \vdots \\ W_{1i}G_{1i}f_{1i}\end{pmatrix} \equiv \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} Q_{Ti}f_{Ti} \\ \vdots \\ Q_{1i}f_{1i}\end{pmatrix}$$

$$\equiv \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \ell_{Ti}(\gamma) \\ \vdots \\ \ell_{1i}(\gamma)\end{pmatrix}. \tag{6.29}$$

Thus, ℓ_i, formed by stacking the elements of the ℓ_{ti}'s, possesses the same structure as (3.8); and $L_N(\gamma) = \frac{1}{N}\sum_{i=1}^{N}\Delta_i f_i$ where (3.9) gives the matrix Δ_i.

6.6.2. *What is the asymptotic distribution of estimators computed using unbalanced data?*

The large-sample distribution of the unbalanced MM estimator $\tilde{\gamma}_w$ depends on the asymptotic properties of the vector

$$\begin{pmatrix} \sqrt{N_T}\frac{1}{N_T}\sum_{i\in\{T\}}\ell^*_{Ti}(\gamma) \\ \vdots \\ \sqrt{N_1}\frac{1}{N_1}\sum_{i\in\{1\}}\ell^*_{1i}(\gamma)\end{pmatrix} = \sqrt{N}L_N(\gamma). \tag{6.30}$$

In sharp contrast to all previous interpretations of the expression $\sqrt{N}L_N(\gamma_0)$, this expression in (6.30) and throughout this subsection merely serves as a notation representing the unequally-normalized vectors specified on the left-hand side of definition (6.30). (So, this expression does not equal the square root of N times L_N listed in (6.28) or (6.29).) The term $\sqrt{N}L_N(\gamma_0)$ defined in (6.30) corresponds to its analogous expression appearing in Section 3 in that it possesses an asymptotic normal distribution with a form comparable to the representations considered above.

In particular, assuming the ℓ^*_{ti} vectors satisfy the same distributional properties maintained for the ℓ_{ti}'s in Section 3, (6.30) converges to a normal distribution possessing the form

$$\sqrt{N}L_N(\gamma_0) \xrightarrow{d} \mathbb{N}\Big(0, \operatorname*{plim}_{N\to\infty}\{\widetilde{V}^*_N\}\Big).$$

The variance–covariance matrix $\widetilde{V}^*_N \equiv V^*_N(\tilde{\gamma}_w)$ has as its (r, s) block the matrix

$$\big\{(r,s)\text{ element of }V^*_N(\gamma)\big\} = \left\{\frac{1}{N_{rs}}\sum_{i\in\{r,s\}}\ell^*_{ri}(\gamma)\ell^*_{si}(\gamma)'\right\}, \tag{6.31}$$

where the notation $\{i \in \{r, s\}\}$ signifies the set of all households with observations in both years r and s, and N_{rs} denotes the total number of households in this set. The approximate large-sample distribution of L_N becomes

$$L_N(\gamma_0) \dot{\sim} \mathbb{N}\left(0, \frac{1}{N}\widetilde{V}^*_N\right),$$

where the (r, s) element of the variance–covariance matrix $\frac{1}{N}V_N^*(\gamma)$ takes the form

$$\left\{(r,s) \text{ element of } \frac{1}{N}V_N^*(\gamma)\right\} = \left\{\frac{1}{\sqrt{N_r N_s}}\frac{1}{N_{rs}}\sum_{i\in\{r,s\}}\ell_{ri}^*(\gamma)\ell_{si}^*(\gamma)'\right\}. \tag{6.32}$$

Similar to the reinterpretation of notation exploited at the beginning of this subsection, the expression $\frac{1}{N}V_N^*$ in (6.32) does not designate the matrix V_N^* divided by N, as has been true in the previous discussion. It merely represents the sample size-normalized variant of the variance–covariance matrix.

Paralleling the steps outlined in Section 3.1, the approximate large-sample distribution of the weighted estimator $\tilde{\gamma}_w$, calculated by minimizing distance function (3.2) using weighted data, is

$$\tilde{\gamma}_w \dot{\sim} \mathbb{N}\left(\gamma_0, \left[\left[\widetilde{S}_N' H_N \widetilde{S}_N\right]^{-1}\left[\widetilde{S}_N' H_N \left(\frac{1}{N}\widetilde{V}_N^*\right) H_N \widetilde{S}_N\right]\left[\widetilde{S}_N' H_N \widetilde{S}_N\right]^{-1}\right]\right), \tag{6.33}$$

where $\widetilde{S}_N \equiv S_N(\tilde{\gamma}_w)$ with

$$S_N(\gamma) = \begin{pmatrix} \frac{1}{N_T}\sum_{i\in\{T\}}\frac{\partial \ell_{Ti}^*}{\partial \gamma'} \\ \vdots \\ \frac{1}{N_1}\sum_{i\in\{1\}}\frac{\partial \ell_{1i}^*}{\partial \gamma'} \end{pmatrix} = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \frac{N}{N_T}\frac{\partial \ell_{Ti}^*}{\partial \gamma'} \\ \vdots \\ \frac{N}{N_1}\frac{\partial \ell_{1i}^*}{\partial \gamma'} \end{pmatrix} = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \frac{\partial \ell_{Ti}}{\partial \gamma'} \\ \vdots \\ \frac{\partial \ell_{1i}}{\partial \gamma'} \end{pmatrix}. \tag{6.34}$$

When implementing estimation procedures with parameters exactly identified – in which case the choice of H_N is irrelevant in the calculation of the estimator – the estimator $\tilde{\gamma}_w$ possesses the simpler asymptotic distribution

$$\tilde{\gamma}_w \dot{\sim} \mathbb{N}\left(\gamma_0, \widetilde{S}_N^{-1}\left(\frac{1}{N}\widetilde{V}_N^*\right)\widetilde{S}_N'^{-1}\right). \tag{6.35}$$

(This distribution is the analog to (3.6).) Finally, the estimator $\hat{\gamma}_w$, computed by minimizing the counterpart to the optimal quadratic-form distance function $C(\gamma)$ given by (3.4), approximately follows the large-sample distribution

$$\hat{\gamma}_w \dot{\sim} \mathbb{N}\left(\gamma_0, \left[\widehat{S}_N'\left(\frac{1}{N}\widetilde{V}_N^*\right)^{-1}\widehat{S}_N\right]^{-1}\right). \tag{6.36}$$

6.6.3. Wrong variance–covariance matrix reported by conventional estimation procedures

Unfortunately, the familiar estimation procedures compute incorrect values for the variance–covariance matrix, even after implementing robust standard errors options. This produces invalid test statistics for some hypotheses. The conventional approaches report variance–covariances based on the following formula to compute $\widetilde{V}_N^*$

$$V_N(\gamma) \equiv \frac{1}{N}\sum_{i=1}^{N}\ell_i(\gamma)\ell_i(\gamma)'.$$

The (r, s) element of this matrix is

$$\left\{(r,s) \text{ element of } \frac{1}{N}V_N(\gamma)\right\} = \left\{\frac{1}{N}\,\frac{1}{N}\sum_{i=1}^{N}\ell_{ri}(\gamma)\ell_{si}(\gamma)'\right\}$$
$$= \left\{\frac{N_{rs}}{N^2}\,\frac{1}{N_{rs}}\sum_{i\in\{r,s\}}\ell_{ri}(\gamma)\ell_{si}(\gamma)'\right\}$$
$$= \left\{\frac{N_{rs}}{N_r N_s}\,\frac{1}{N_{rs}}\sum_{i\in\{r,s\}}\ell^*_{ri}(\gamma)\ell^*_{si}(\gamma)'\right\}. \tag{6.37}$$

Comparing (6.32) to (6.37) reveals discrepancies in the off-diagonal elements of the valid specification of $\frac{1}{N}V_N^*$ and the reported value $\frac{1}{N}V_N$. The relationship between the (r, s) elements of these matrices is

$$\left\{(r,s) \text{ element of } \frac{1}{N}V_N^*(\gamma)\right\} = \frac{\sqrt{N_r N_s}}{N_{rs}}\left\{(r,s) \text{ element of } \frac{1}{N}V_N(\gamma)\right\}. \tag{6.38}$$

This formula shows how to adjust each element of the reported matrix – the right-hand side elements in (6.38) – to the appropriate values.

Inspection of (6.38) reveals that the diagonal blocks of these matrices are the same since, with $r = s$, $N_{rs} = N_r = N_s$. This implies that standard errors, t-statistics, and test statistics reported by the regular estimation procedures are valid as long as no constraints are imposed across equations. When restrictions are considered across equations, the off diagonal blocks come into play and the elements in these blocks differ depending on the relative sample sizes in distinct periods. In many longitudinal datasets, observations are dropped after the period they first attrit from the panel. Consequently, assuming period r comes before period s and some attrition occurs, then $N_r > N_s$. If all households present in period s were also observed in period r, then $N_{rs} = N_s$. In this case the conversion factor becomes

$$\frac{\sqrt{N_r N_s}}{N_{rs}} = \frac{\sqrt{N_r}}{N_s}. \tag{6.39}$$

Thus, the (r, s) element reported by conventional estimation is too low, and it must be enlarged by the ratio of the square roots of the early sample size to the later one to calculate the correct covariance. So, if the period r sample is twice as large as the period s sample, the covariances associated with coefficients across the year r and s equations must be multiplied by $\frac{\sqrt{2}}{\sqrt{1}} \approx 1.4$. Of course, if all households are observed in all years, then $N_{rs} = N_r = N_s = N$, and no adjustments are required.

6.7. *Weighting and unbalanced data in the estimation of quantile specifications*

The above procedures readily apply in estimating the parameters of conditional quantile relationships using a stratified and/or an unbalanced sample. When faced with a stratified sample in a longitudinal data context, selecting instrumental variables according to

(6.24) in the implementation of Section 5.1 quantile estimation approach yields consistent estimates of the coefficients $\rho_{t1}, \ldots, \rho_{tr}, \beta_t$. As noted in Section 6.5, the selection of weights w_{ti} in (6.24) depends on precisely which population a dataset's weights emulate, and which population a researcher wishes to replicate. Often an analyst estimates intertemporal relationships using balanced data, restricting the sample to include observations for periods 1 through T for all individuals i who remain sample members during this time horizon. In such a situation, one would select $w_{ti} = w_{Ti}$ for all t when forming the weighted instrumental variables in (6.24).

When estimating dynamic quantile specifications using unbalanced data, the discussion of Section 6.6 applies fully. One carries out nonlinear instrumental variable estimation using weights specified by (6.26) and the formulation of L_N given by (6.28) and (6.29). This formulation applies directly when implementing the multi-equation method described above with f_{ti} specified by (5.11). This framework also permits implementation of the nonlinear 2SLS procedure discussed above, which assumes the structural errors U_{ti} are distributed independently over time as well as across individuals. In this 2SLS case, f_{ti} in (6.28) is a scalar and the specification of the H_N matrix in the formulation of the distance function (3.2) takes the form:

$$H_N = \left[\iota' \otimes \left[\frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} Q_{ti} Q_{ti}' \right]^{-1} \right],$$

where ι in this Kronecker product refers to a column vector of 1's of dimension T, and Equations (6.29) define the instrumental variables Q_{ti}.

7. An empirical application to wage dynamics

This section introduces a set of empirical examples to illustrate the econometric methods presented in the previous four sections. These examples characterize the dynamic properties of hourly wages of men during the period 1980–1991 using the Panel Study of Income Dynamics (PSID). This empirical analysis is not intended to provide a comprehensive investigation of wage dynamics; instead, its aim is to enhance the accessibility of the procedures discussed in this chapter to practitioners. The section reports empirical findings applying many of these procedures, explaining implementation in a concrete context and comparing standard-error estimates obtained through the use of both classical and bootstrap approaches. It also highlights the differences in estimation from using balanced data, where individuals are restricted to have data for every period t, versus unbalanced data, where individuals who are only available for part of the sample period are retained. After illustrating the proper use of stratified sample weights, the section ends by using quantile regression procedures to characterize the dynamic properties of median hourly wages for men.

Section 7.1 summarizes the data used, while Section 7.2 estimates covariograms using the method described in Section 4.4. Section 7.3 uses the methods of Sections 4.5

and 4.7 to illustrate how the autoregressive parameters can be estimated alone, or jointly along with the moving-average parameters. Section 7.4 reports the results of the estimation proposed in Section 7.3. Section 7.5 provides bootstrapped standard errors for comparison to the asymptotic standard errors reported in Section 7.4. Section 7.6 applies results in Section 6.6 to illustrate the utilization of information from unbalanced data, and Section 7.7 applies the results in Section 6.4 to show the correct use of stratified sampling weights. Section 7.8 is an application of quantile regressions to estimate ARMA processes, as discussed in Section 5. Finally, Section 7.9 summarizes the findings.

7.1. Data description and prototype model

Data are drawn from the randomly-designed sample of the Michigan Panel Study of Income Dynamics.[39] The dataset consists of 959 observations on prime-age males for the years 1980–1990, a total of 11 years of data for each individual. Only males 25–46 years old in 1980 are included in the sample. The wage rate is defined to be annual real earnings deflated by the Consumer Price Index and divided by hours of work reported during the year.

Because the PSID is a stratified sample, one must apply weights to adjust for differences between the sample and national populations in order to obtain consistent estimates of population parameters. Regrettably, most standard software routines do not use weights properly in calculating standard errors when weights adjust for stratified sampling; these routines instead assume weights merely adjust for heteroscedasticity. Routines that do make proper adjustments are often termed "survey sampling" procedures. Given the illustrative nature of the empirical examples presented below, and the fact that most economists ignore weighting, the following empirical examples do not use weighting in calculating estimates. To assess the impact of weighting, Section 7.7 presents the findings of several exercises.

The following exercises assume that wages obey the regression/median equation:

$$\omega_{ti} = X_{ti}\beta + U_{ti}, \tag{7.1}$$

where ω_{ti} measures the growth in an individual's hourly wages from period $(t-1)$ to t. This equation is a special case of relation (2.10), with π and ψ set equal to zero and β constrained across time.[40] The majority of the analysis presumes that the disturbances in (7.1) follow the ARMA process specified in Equation (2.3)

$$a(L)U_{ti} = m(L)\varepsilon_{ti},$$

[39] The panel study's procedures and methods are detailed in Hill (1992).

[40] Regression coefficients are constrained to be equal across years. Formal hypothesis tests easily accept this restriction at conventional levels of significance.

which the following empirical applications respecify as

$$U_{ti} = \sum_{j=1}^{p} a_{jt} U_{(t-j)i} + \sum_{j=0}^{q} m_{jt} \varepsilon_{(t-j)i}. \tag{7.2}$$

(Note, the coefficients a_{jt} in (7.2) have been redefined to have the opposite sign as the corresponding coefficients appearing in relation (2.3) and the previous discussion; expression (7.2) takes the form typically found in empirical earnings literature.) Both mean and quantile regression techniques can be used to consistently estimate the parameters of interest. For simplicity, the subscript i will be omitted from equations when this causes no confusion.

Least squares estimation of the reduced-form model (7.1) produces the residuals $\widehat{U}_t$ used in much of the subsequent empirical exercises. The variables incorporated as regressors in X_t are designed to capture the measured component of wage growth, including four education dummy variables, a quadratic in age, full interactions between the age polynomial and the education dummy variables and a dummy variable for each year of the sample after the first. With time effects included in X_t, the errors U_t represent the vector of deviations in individuals' wage growth from averages in the population in period t, after accounting for age and education. Thus this exercise seeks to characterize the correlation across years in unmeasured (residual) wage growth.

7.2. Estimation of autocorrelations

Prior to estimating specifications of model (7.1), it is necessary to investigate two questions: (1) Do the vectors $U_1, \ldots, U_T$ satisfy the weak stationarity property implicitly assumed in the multiple time series specifications presented in Equation (7.2)? (2) What are the orders of the autoregressive and the moving average lag polynomials in the multivariate ARMA process that best describe the intertemporal variation in the U_t's?

7.2.1. Estimating covariograms

Estimating autocovariances provides the essential information needed to answer these questions. Section 4.4 describes the approach utilized to calculate these quantities implementing the joint generalized least squares or seemingly unrelated regression framework specified by Equation (4.24) using the fitted values of U_t, $\widehat{U}_t$. This procedure estimates the second moments $E(U_t U'_{t-k})$ for each value of t and k, setting each moment equal to its sample analog $\widehat{U}_t \widehat{U}_{t-k}$, which consistently estimates population autocovariances when averaged. Weak stationarity follows if one can accept the hypothesis that these second moments are independent of t for any k (i.e., independent of year for any lag). One can test weak stationarity with a standard F-test on the joint hypothesis that for a given k, $E(U_t U_{t-k})$ is the same for all t. Further, the properties of these estimated second moments provide information allowing one to choose which ARMA process best fits the data.

The estimation of the sample moments is done within a seemingly unrelated regression context in order to account for unspecified heteroscedasticity and autocorrelation across time for a given individual. Thus the values $\widehat{U}_t\widehat{U}_{t-k}$ for all available combinations of t and k are constructed for each individual to form the system of Equations (4.24).

7.2.2. *Implications of covariograms for stationarity and ARMA specifications*

Table 1 presents estimates of the covariogram associated with specification (7.1). The first row presents estimates for autocovariances of order $k = 0, 1, \ldots, 6$ when the autocovariances of the specified order are constrained to be equal across time. These are the constrained estimates of θ from the version of Equation (4.24) described above. The second row lists the autocorrelation coefficients implied by these estimates. The third row reports the minimum and maximum estimates of the autocovariances when the θ's are not restricted to be equal over time. The fourth row reports the test results for the hypothesis that the kth order autocovariance is equal across periods, e.g., the hypothesis is that the covariance between U_{1981} and U_{1982} is the same as the covariance between U_{1982} and U_{1983}.

Table 1 provides answers to the two questions posed above. The F-test in row 4 is easily accepted for all k. Therefore, the data do accept a weak stationarity restriction.

The second question, how to model the error structure process in a tractable way, amounts to picking an ARMA process that best fits the data. Autoregressive processes lead to autocovariances that, at orders higher than the order of the process, gradually fall to zero. Moving-average processes exhibit autocovariances that sharply drop to zero once one moves to an order higher than the order of the moving average. These two theoretical predictions are the guidelines from which to specify the error structure. Note that these predictions involve evaluating magnitudes (i.e., absolute values) of the coefficients.

Using the above theoretical predictions as guidelines, it is easy to see the existence of a first-order moving average in the data. The first-order autocovariance term, estimated to be -0.048, is by far the largest in magnitude. The second-order term, while still statistically different from zero, takes a sharp drop to -0.006, an eighth the size of the first. The higher order autocovariance terms get progressively smaller and are all statistically indistinguishable from zero.

The sharp drop after the first-order autocovariance term suggests a first-order moving-average process, but it also suggests a short autoregressive process. A lengthy autoregressive process would not have autocovariance terms that drop off so fast. The gradual fall in the terms of order two and higher invites one to consider a low-order autoregressive process. Based on the autocovariance coefficients and prior work, this section will investigate models with a first-order moving average component and either a first- or second-order autoregressive process (ARMA(1, 1) or ARMA(2, 1)). One could investigate a wider class of specifications, but these two specifications should serve to sufficiently illustrate the methods without getting bogged down in repetitive tables. For

Table 1
Covariogram for wage growth residuals

Statistics	Lag (order)						
	0	1	2	3	4	5	6
Constrained autocovariances (standard errors in parentheses)	0.131 (0.007)	−0.048 (0.004)	−0.006 (0.002)	−0.004 (0.0025)	−0.003 (0.003)	−0.001 (0.003)	0.002 (0.003)
Autocorrelation	1	−0.37	−0.05	−0.03	−0.02	−0.01	0.01
Range of autocovariances	(0.120, 0.171)	(−0.035, 0.067)	(−0.011, 0.005)	(−0.021, 0.002)	(−0.013, 0.005)	(−0.004, 0.007)	(−0.006, 0.005)
Test for constant autocovariance (probability of event in parentheses)	yes (0.46)	yes (0.60)	yes (0.63)	yes (0.32)	yes (0.71)	yes (0.85)	yes (0.80)

Note. Based on seemingly unrelated regression model.

notational simplicity, equations will be in terms of the ARMA(2, 1) specification, and the changes that are required to estimate the ARMA(1, 1) process will be noted.

7.3. Empirical specifications for ARMA error process

This section describes specifications and methods used to estimate the autoregressive and moving-average coefficients in Equation (7.2). The estimation done here uses the framework outlined in Section 4 with $\pi = 0$ and $\psi = 0$, meaning that there are no right hand endogenous variables. Combining Equations (7.1) and (7.2), the researcher wishes to estimate the parameters of the following equation:

$$\omega_t - X_t\beta + a_1(\omega_{t-1} - X_{t-1}\beta) + a_2(\omega_{t-2} - X_{t-2}\beta) = \varepsilon_t + m_1\varepsilon_{t-1}, \tag{7.3}$$

an ARMA(2, 1) specification. When $a_2 = 0$, Equation (7.3) specifies an ARMA(1, 1) process.

7.3.1. Specifications for estimating only autoregressive coefficients

Section 4.5 outlines methods for directly estimating the autoregressive parameters of ARMA processes. Starting with Equation (7.1) one can follow Section 4.5 to estimate both the structural parameters β and the autoregressive parameters of the error process. Assuming that the moving-average process is first-order implies that error terms from two or more periods back are predetermined and can be used as instruments. In this case, define a system of equations of the following form (modeled after Equations (4.10) and (4.27)):

$$\begin{aligned} f_t^{(8)} &= \omega_t - X_t\beta, \\ f_{tk}^{(9)} &= [\omega_{t-k} - X_{t-k}\beta] \\ &\quad \times \left[(\omega_t - X_t\beta) - a_1(\omega_{t-1} - X_{t-1}\beta) - a_2(\omega_{t-2} - X_{t-2}\beta)\right] \\ k &= 2, \ldots, (t - 1983),\ t = 1983, \ldots, 1990. \end{aligned} \tag{7.4}$$

Stacking these equations to obtain $f_i(\beta, a) = (f_t^{(8)\prime}, f_{tk}^{(9)\prime})'$ for all t creates a model in the form of Equation (3.7). This can be estimated using nonlinear three-stage least squares. Following the discussion in the second half of Section 4.5, if one wishes to estimate only the autoregressive parameters then one could use a consistent estimate of β to form

$$\begin{aligned} f_{tk}^{(10)} &= \widehat{U}_{t-k} \times (\widehat{U}_t - a_1\widehat{U}_{t-1} - a_2\widehat{U}_{t-2}), \\ k &= 2, \ldots, (t - 1983),\ t = 1983, \ldots, 1990, \end{aligned} \tag{7.5}$$

where $\widehat{U}_t = \omega_t - X_t\hat{\beta}$ and $\hat{\beta}$ is the ordinary least squares estimate from OLS in Equation (7.1). This equation is an ARMA(2, 1) version of Equation (4.30). The second half of the product in $f_{tk}^{(10)}$ is the g_t of Equation (3.18); it is multiplied by an orthogonal regressor in order to provide the identifying restrictions.

Equation (7.5) is the specification estimated, with one modification. Instead of estimating the model with all available predetermined variables as instruments, i.e., all residuals from two or more periods back, this section uses a linear combination of those predetermined variables – specifically the linear combination that brings them closest to being the optimal instruments discussed in Sections 3.2.2 and 4.1.4: namely, $E(\frac{\partial f_{ti}}{\partial a_1}|X_{ti})$ and $E(\frac{\partial f_{ti}}{\partial a_2}|X_{ti})$. This is done by regressing $\widehat{U}_{t-1}$ and $\widehat{U}_{t-2}$ on all available predetermined and exogenous variables. This yields predicted values for these two quantities for every year t, which can then be used as instruments. Formally, let $\widehat{U}_t^p$ be the predicted value of $\widehat{U}_t$ based on a regression on *all* its previous (predetermined) lags. For example, regress $\widehat{U}_{1984}$ on $\widehat{U}_{1981}$, $\widehat{U}_{1982}$, and $\widehat{U}_{1983}$. Then use the predicted value, $\widehat{U}_{1984}^p$, as an instrument in Equation (7.5) where $t = 1985$. In the 1986 equation, it will also serve as an instrument, but since it is two periods back it is predetermined in the 1986 equation and therefore can be perfectly predicted by itself. This allows one to get instruments that are close to the optimal instruments but are uncorrelated with the first-order moving-average error term.

Implementing the above specification yields

$$\begin{aligned} f_t^{(11)} &= \widehat{U}_{t-1}^p \times (\widehat{U}_t - a_1\widehat{U}_{t-1} - a_2\widehat{U}_{t-2}) = \widehat{U}_{t-1}^p \times g_t, \\ f_t^{(12)} &= \widehat{U}_{t-2}^p \times (\widehat{U}_t - a_1\widehat{U}_{t-1} - a_2\widehat{U}_{t-2}) = \widehat{U}_{t-2}^p \times g_t, \\ & t = 1983, \ldots, 1990. \end{aligned} \tag{7.6}$$

Stacking these equations to obtain $f_i(a_1, a_2) = (f_t^{(11)\prime}, f_t^{(12)\prime})'$ for $t = 1983, \ldots, 1990$ creates a model of the form of (3.7). Using these two projected instruments is convenient because they are sufficient to identify both a_1 and a_2, while reducing the number of equations in the system makes it easier to estimate computationally and decreases programming time. In the case of the ARMA(1, 1), $f_t^{(12)}$ is dropped from estimation as $f_t^{(11)}$ is sufficient to identify the system.

7.3.2. *Specifications for estimating autoregressive and moving-average coefficients jointly*

Since the covariograms gave strong evidence of a moving-average process, one might wish to estimate jointly the autoregressive and moving-average parameters as described in Section 4.7. This can be done by adding the relevant moment restrictions to the estimation. Using the notation from above, one would estimate:

$$\begin{aligned} f_{tk}^{(13)} &= \widehat{U}_{t-k} \times (\widehat{U}_t - a_1\widehat{U}_{t-1} - a_2\widehat{U}_{t-2}) = \widehat{U}_{t-k} \times g_t, \\ f_t^{(14)} &= (\widehat{U}_t - a_1\widehat{U}_{t-1} - a_2\widehat{U}_{t-2})^2 - \sigma_{11} = g_t^2 - \sigma_{11}, \\ f_t^{(15)} &= (\widehat{U}_t - a_1\widehat{U}_{t-1} - a_2\widehat{U}_{t-2})(\widehat{U}_{t-1} - a_1\widehat{U}_{t-2} - a_2\widehat{U}_{t-3}) - \sigma_{12} \\ &= g_t \times g_{t-1} - \sigma_{12}, \\ & k = 2, \ldots, (t - 1983),\ t = 1983, \ldots, 1990. \end{aligned} \tag{7.7}$$

The value σ_{11} in $f_t^{(14)}$ estimates the variance of the residual term $(\varepsilon_t + m_1\varepsilon_{t-1})$ while σ_{12} in $f_t^{(15)}$ estimates the residual first-order autocovariance. To estimate the moving average component, use the estimates of the variance and the first-order autocovariance to derive the parameter m_1. This requires extracting it from the equations $\sigma_{11} = (1 + m_1^2)\sigma_\varepsilon^2$ and $\sigma_{12} = m_1\sigma_\varepsilon^2$, which follow from standard time-series results.[41]

The estimation of Equation (7.7) can be simplified through the use of near-optimal instruments as discussed above. Using similar notation, the optimal instruments are predicted using all predetermined variables, and then these estimated instruments are introduced into the estimation, yielding:

$$
\begin{aligned}
f_t^{(11)} &= \widehat{U}_{t-1}^p \times (\widehat{U}_t - a_1\widehat{U}_{t-1} - a_2\widehat{U}_{t-2}) = \widehat{U}_{t-1}^p \times g_t,\\
f_t^{(12)} &= \widehat{U}_{t-2}^p \times (\widehat{U}_t - a_1\widehat{U}_{t-1} - a_2\widehat{U}_{t-2}) = \widehat{U}_{t-2}^p \times g_t,\\
f_t^{(14)} &= (\widehat{U}_t - a_1\widehat{U}_{t-1} - a_2\widehat{U}_{t-2})^2 - \sigma_{11} = g_t^2 - \sigma_{11},\\
f_t^{(15)} &= (\widehat{U}_t - a_1\widehat{U}_{t-1} - a_2\widehat{U}_{t-2})(\widehat{U}_{t-1} - a_1\widehat{U}_{t-2} - a_2\widehat{U}_{t-3}) - \sigma_{12}\\
&= g_t \times g_{t-1} - \sigma_{12},\\
t &= 1983, \ldots, 1990.
\end{aligned}
\tag{7.8}
$$

Stacking these equations to obtain $f_i(a_1, a_2, \sigma_{11}, \sigma_{12}) = (f_t^{(11)\prime}, f_t^{(12)\prime}, f_t^{(14)\prime}, f_t^{(15)\prime})'$ for $t = 1983, \ldots, 1990$ creates a model of the form of (3.7). When the model tested is ARMA(1, 1), the a_2 parameter is set to zero and the second moment restriction, $f_t^{(12)}$, is unnecessary for identification and is dropped.

7.3.3. Estimators for ARMA coefficients

This estimation can be done in several different ways, two of which are mentioned here. The first is to use a method of moments procedure to relate the ARMA parameters to the sample mean and autocovariances through the system of moment restrictions as given in Equations (7.6) and (7.8). The equations are estimated with a constant as the only instrument specified for the software. This gives consistent results with asymptotically valid standard errors and covariances.

The second method is similar; the researcher uses nonlinear three-stage least squares with a constant as the only designated instrument. This also produces asymptotically valid results. Note that the "real" instruments are embedded in the equations to be estimated and differ from equation to equation. Thus the researcher instruments manually instead of using software commands to designate an instrument set. Reported below are estimates from the method-of-moments approach.

[41] The quadratic in the equations yields two answers. The convention is to use the root that is less than one in absolute value, making the series invertible.

7.4. Empirical findings for ARMA estimation

Turning to the results, Tables 2 and 3 present estimates for the two candidate ARMA specifications generating the individual error terms in wage growth. Results are unweighted and standard errors are given in parentheses. This subsection will discuss each table in turn. The goals are to compare the specifications to see which better fits the data and to determine how robust the results are across models.

7.4.1. Estimates of only autoregressive coefficients

Table 2 presents the coefficient estimates for the autoregressive component of the ARMA(1, 1) and ARMA(2, 1) models specified by Equation (7.6). Recall that Equation (7.6) estimates only the autoregressive parameters of each ARMA model. Looking at the ARMA(1, 1) specification, estimating just a_1 gives a smaller coefficient in magnitude than when a_2 is not constrained to be zero, -0.134 vs. -0.189. So from year to year approximately 13 to 19 percent of the residual variation in wage growth is undone the following year through the autoregressive parameter. In terms of their sampling dis-

Table 2
Estimates of only autoregressive coefficients for wage growth error structure

ARMA(p, q)	a_1	a_2
(1, 1)	−0.134	
	(0.0354)	
(2, 1)	−0.189	−0.024
	(0.0491)	(0.0254)

Note. Specification based on Equation (7.6) (standard errors in parentheses).

Table 3
Joint estimates of autoregressive and covariance parameters for wage growth error structure

ARMA(p, q)	Direct estimates				Implied values from delta method	
	a_1	a_2	σ_{11}	σ_{12}	σ	m_1
(1, 1)	−0.151		0.131	−0.060	0.305	−0.645
	(0.0339)		(0.0074)	(0.0040)	(0.0076)	(0.0279)
(2, 1)	−0.186	−0.020	0.132	−0.062	0.300	−0.686
	(0.0405)	(0.0225)	(0.0089)	(0.0052)	(0.0075)	(0.0441)

Note. Specification based on Equation (7.8) (standard errors in parentheses).

tributions, the two estimates are several standard deviations from 0 and so the effect is significantly different from zero.

The ARMA(2, 1) specification reports an estimate for the a_2 parameter. The estimate is negative and small. The a_2 estimate, -0.024, cannot be statistically distinguished from zero, despite the fact that the standard errors are tighter than those reported on the a_1 coefficient. Regardless, the value indicates that if there is a second-order term it is probably negative and small.

7.4.2. *Estimates of autoregressive and moving-average coefficients jointly*

Table 3 reports the results from the joint estimation of the autoregressive and moving-average coefficients as specified by Equation (7.8) in Section 7.3.2. The first four columns are self-explanatory in that the parameters listed appear directly in Equation (7.8); the last two columns are delta method extrapolations of moving-average parameters. These columns give both the implied standard deviation of the white noise process, σ, and the coefficient on the first-order moving average, m_1.

The results are similar, in some respects, to what was observed in Table 2. The first-order autoregressive parameter a_1 is reported as -0.151 in the ARMA(1, 1) and -0.186 in the ARMA(2, 1). These estimates are close to one another and close to the estimates found when just the autoregressive parameters were estimated. The a_2 parameter in the ARMA(2, 1) model looks almost exactly as it did when only the autoregressive parameters were estimated. It remains small and statistically indistinguishable from zero.

Turning to the covariance terms, the table shows that both specifications return stable and precisely estimated results. The σ_{11} coefficient is about 0.13 and the σ_{12} coefficient is almost half as large at about -0.06. The standard errors on these estimates are small, providing confidence of a reasonably good estimate. Notably, the σ_{12} estimate is clearly not zero, reinforcing the hypothesis that there is a first-order moving-average process. From these two parameter estimates one can get the implied estimates for the standard deviation on the white noise process, σ, and the coefficient on the moving average component, m_1. Using linear extrapolation (the delta method) one can also compute an asymptotic approximation for standard errors. The reported coefficient on σ is about 0.3 in both specifications. For m_1, the two specifications return values of -0.645 and -0.686, respectively. The standard error is less than 0.03 in the ARMA(1, 1) but 0.044 in the second. Regardless, both procedures find evidence of what previous analysis of the covariogram indicated – a large negative first-order moving average.

In summary, the data strongly support the hypothesis of serial correlation in the error terms. The first-order autoregressive component is somewhere between -0.13 and -0.19. The moving-average parameter is between -0.6 and -0.7. The second-order autoregressive lag is probably slightly negative but cannot be distinguished from zero. Looking back to the covariogram in Table 1, these parameter estimates easily account for the large negative correlation between wage residuals in adjacent years. They also fit with the observed lack of correlation beyond the first lag.

7.5. Bootstrapping ARMA models using panel data

This discussion very briefly presents findings illustrating the consequences of utilizing bootstrap procedures to fit the two models considered above. The goal is to compare the standard errors calculated by classical first-order normal asymptotic theory and those computed by the residual resampling method. This section illustrates the bootstrapped version of the method-of-moments estimator considered above.

7.5.1. Estimates with bootstrapped standard errors

There may be some doubt as to the effectiveness of using first-order asymptotic theory as a guide to constructing standard errors. As always, the asymptotic theory offers only a guideline. At times there is reason to question the validity of this guideline. This may be because the model was misspecified or because the sample is too small to be well-approximated by its large-sample distribution. One can cross-check the validity of the asymptotic approximation by comparing it to a bootstrapped estimate. Instead of relying on a linear extrapolation at infinite sample size to guide standard error calculation, one assumes that the sampled data is representative of the population density, and thus can be used to evaluate the sampling distribution of the estimators.

No special techniques are required to produce these results. Bootstrapping begins with the residuals first described in Section 7.2.1 and used in Section 7.4 to estimate the ARMA processes. Each individual's set of residuals is given an equal probability of being drawn with replacement. To preserve the serial correlation, the resampling is done over the individuals, not over each year of each individual. The sample consists of 959 individuals, so 959 draws are made in each resample. 1000 resamples are performed and Equations (7.6) and (7.8) are estimated, for both the ARMA(1, 1) and the ARMA(2, 1) specifications. These 1000 estimated coefficients are used to create a sampling distribution for each estimator. The standard deviation of these 1000 coefficients is the standard error for the estimator. The following subsection compares the asymptotic standard errors reported in Tables 2 and 3 to those implied by the bootstrap procedure. Once again, the estimation is performed on the unweighted sample.

7.5.2. Implications of bootstrap estimates

Table 4 reports comparisons for specifications of Equations (7.6) and (7.8) estimated in Section 7.4. It first estimates the subset of autoregressive parameters as in Equation (7.6), and then adds the covariance parameters as in Equation (7.8). Each specification lists the previously developed asymptotic estimation results and is immediately followed by the bootstrap results.

All of the standard errors are wider for the bootstrapped sample, typically – although not always – on the order of 25 percent. Thus the asymptotic standard errors are biased downwards for this sample size. Turning to parameter estimates, when the autocovariance terms are estimated alone, the bootstrap gives results that are very close to the

Table 4
Bootstrapped estimates of ARMA processes for wage growth error structure

ARMA(p, q)	Estimation method	a_1	a_2	σ_{11}	σ_{12}
(1, 1)	Autoregressive subset estimation	−0.134			
	Asymptotic theory	(0.0354)			
	Autoregressive subset estimation	−0.136			
	Bootstrap with 1000 replications	(0.0406)			
	Full estimation	−0.151		0.131	−0.060
	Asymptotic theory	(0.0339)		(0.0074)	(0.0040)
	Full estimation	−0.115		0.120	−0.054
	Bootstrap with 1000 replications	(0.0448)		(0.0088)	(0.0048)
(2, 1)	Autoregressive subset estimation	−0.189	−0.024		
	Asymptotic theory	(0.0491)	(0.0254)		
	Autoregressive subset estimation	−0.173	−0.017		
	Bootstrap with 1000 replications	(0.0540)	(0.0274)		
	Full estimation	−0.186	−0.020	0.132	−0.062
	Asymptotic theory	(0.0405)	(0.0225)	(0.0089)	(0.0052)
	Full estimation	−0.117	0.003	0.116	−0.052
	Bootstrap with 1000 replications	(0.0807)	(0.0331)	(0.0123)	(0.0076)

Note. Specifications based on Equations (7.6) and (7.8) (standard errors in parentheses).

original estimates. When the joint system is estimated, the bootstrapped a_1 and a_2 parameter estimates move noticeably. The largest effect is for the a_1 parameter estimate in the ARMA(2, 1) model. It goes from −0.186 in the method-of-moments routine to a bootstrapped estimate of −0.117.[42] Unsurprisingly, the standard error doubles. Note that the bootstrapped estimate of a_1 is still close to the original range of −0.13 to −0.19 observed in the original data, but it has jumped from the high end to the low end. The estimates for σ_{11} and σ_{12}, on the other hand, are fairly stable.[43]

7.6. Results based on balanced versus unbalanced data

Practitioners of econometrics often are confronted with thorny problems stemming from data collection. Although these problems do not draw as much attention in the literature

[42] These large changes are possible when one is dealing with a nonlinear estimation method such as the one used here. But the magnitude of the change is cause for wonder. Preliminary data research uncovered several large outliers, which are not unusual in sample data, especially when they have been first-differenced. This leads to an extremely fat-tailed distribution for the residuals. This could make the 4-equation system (7.8) vulnerable to outliers due to its squared term. These vulnerabilities were apparently uncovered in the bootstrap estimation.

[43] For a comprehensive and enlightening discussion of the issues involved in bootstrap estimation, see the Handbook chapter by Horowitz (2001).

as other, more provocative subjects, they can have a large effect on the estimates' validity. This subsection and the next deal with a brief application of the two data issues discussed in Section 6: unbalanced data and stratified sampling.

The first concern, unbalanced data, stems from the fact that many sampled units (in this case, people) do not have data in one or more years of a panel dataset. One could assume random attrition (which probably is not true but is very convenient) and form estimates using just those people having all observations.[44] This ignores potentially valuable information. In the PSID sample utilized thus far, roughly a third of the observations were dropped due to incomplete data over the sample period. Section 7.6 looks at how one might conveniently estimate the covariogram for the PSID data using an unbalanced sample.

Section 7.7 looks at stratified sampling weights. Datasets frequently oversample certain groups in order to provide researchers with a more detailed picture of a small segment of the population. This oversampling destroys the randomness of the sample and, as noted in Section 6, requires some care in correcting. Section 7.7 will give an example correctly using the stratified sampling weights by designating them as instruments.

7.6.1. Estimates with unbalanced data

Researchers prefer to use all the data available to conduct an empirical analysis to enhance the efficiency of estimation. For this reason, many would like to recapture the information contained in observations that do not have data for one or more years of the sample.

The seemingly unrelated regression method used in Section 7.2 requires complete data, i.e., the dataset must be balanced. This has the unfortunate result that one must throw out observations that contain all the necessary information but are missing even just one year. Section 6.6 details a method to recover this lost information while staying within the simple seemingly unrelated regression framework. Applying this method assigns a zero person-year weight to those person-year observations that are missing and proceeds with estimation as if the data were balanced. This amounts to replacing the entire equation for the person-year by zero. In conducting this estimation, the procedure formulates weights – or adjusts stratified sample weights if they are used in the analysis – for the remaining observed data to reflect the number of missing observations associated with each equation.

As discussed in Section 6.6.3, an additional correction is necessary to perform multivariate, multi-equation tests when using unbalanced data. The off-diagonal elements of the sample covariance matrix require the correct denominator. Working from the

[44] Fitzgerald, Gottschalk and Moffitt (1998), in an analysis of PSID attrition, conclude that use of the available weights maintains the representativeness of the survey.

formulas already used to correct the variance terms,[45] multiply the i, j element of the cross-equation covariance matrix by the square roots of the number of available observations in equation i and the number of available observations in equation j, then divide by the number of common observations between i and j. This corrected covariance matrix can then be used to test the weak stationarity hypothesis first discussed in Section 7.2.

7.6.2. *Implications of estimating with unbalanced data*

Table 5 presents estimates of the covariogram when using the unbalanced data. Its format is the same as Table 1. The zero-order variance term is larger at 0.162. The higher order terms also tend to be larger so that the autocorrelation terms are very close to what they were in Table 1. Instead of the first-order autocorrelation of -0.37, the unbalanced sample reports the slightly weaker -0.35. The second order term was -0.05 in the balanced data but now is -0.04. Regardless, the inferences made in Section 7.2 are unchanged. The tests of weak stationarity all fail to reject at a 5 percent significance level, if only barely. Thus, there is room to doubt the assumption, but not nearly enough evidence to overturn it.

7.7. *Results based on weighted versus unweighted data*

Up to now this section has ignored the implications of weighting stratified samples like the PSID. Section 6 explained some of the difficulties that can arise in properly weighting stratified samples. In this subsection there is a short discussion of how the principles outlined in Section 6 apply to the estimation done here.

7.7.1. *Estimates with stratified sample weights*

Several statistical packages offer a weighting option in their estimation routine. Usually this is the standard weighted least squares procedure designed to account for heteroscedasticity. If a researcher uses weighted least squares to calculate standard errors, the package premultiplies the regressor and the regressand by the square root of the weight and then proceeds with the rest of the estimation as usual. In the linear case it reports estimator variances of the form $\hat{\sigma}^2(X'\Omega X)^{-1}$ where $\hat{\sigma}^2 = \frac{1}{N}\sum_1^N \hat{\varepsilon}^2$ and Ω is a diagonal matrix with the weight w_i for each observation on the diagonal. This is correct if the weighting is designed to correct for heteroscedasticity of the form $\frac{\sigma^2}{w_i}$. It is not correct for the case of a stratified sample. The general framework is outlined in Section 6. The proposed solution when using instrumental variables techniques is to premultiply the instruments by the weight w_i (making sure that the weight is properly normalized).

[45] Recall that since each equation estimates a single parameter, the mean, the cross-equation covariance system is exactly the same size as the unrestricted coefficient covariance matrix.

Table 5
Covariogram based on unbalanced wage growth residuals

Statistics	Lag (order)						
	0	1	2	3	4	5	6
Constrained autocovariances	0.162	−0.056	−0.007	−0.004	−0.004	−0.001	0.003
(standard errors in parentheses)	(0.008)	(0.004)	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)
Autocorrelation	1	−0.35	−0.04	−0.02	−0.02	−0.01	0.02
Range of autocovariances	(0.148, 0.206)	(−0.080, −0.043)	(−0.020, 0.004)	(−0.026, 0.005)	(−0.012, 0.008)	(−0.011, 0.003)	(0.002, 0.003)
Test for constant autocovariance	yes	yes	yes	yes	yes	yes	yes
(probability of event in parentheses)	(0.40)	(0.47)	(0.46)	(0.06)	(0.10)	(0.73)	(0.58)

Note. Based on seemingly unrelated regression model.

When one is estimating with the method outlined in Sections 7.3 and 7.4 (where the only instrument is a constant) one could simply designate the stratified sampling weights as the instruments. These new instruments will impose the correct weighting for consistent estimates and will yield asymptotically correct standard errors. Before giving results there are two things which should be noted. First, as noted in Equation (6.7), stratified weights in survey data are often designed to sum to the size of the total population. These w_i^* weights can be converted to w_i weights by dividing them by their average value. This is the strategy followed in Section 6.3. Thus all weights discussed in this subsection are normalized weights. Second, as explained in Section 6.5, the correct weight to use for all years of the PSID is the weight given in the last year of the sample, as this weight reflects the longitudinal changes in the data.

7.7.2. Implications of stratified sampling weights

Tables 6 and 7 redo the estimation of Sections 7.2 and 7.4 using the balanced data and sample weights from 1990. Table 6 is the covariogram and is designed the same as Table 1. The methodological difference is that in the seemingly unrelated regression framework used to construct the covariogram, the dependent variable was premultiplied by the weight for 1990.[46] One could obtain the same result by switching to a three-stage least squares framework and designating the weight as the only instrument.

The results are quantitatively different but the implications are identical. The variance (zero-order autocovariance) falls from an unweighted value of 0.13 to a weighted value of 0.11. All the other terms tend to be proportionately lower and so the autocorrelations are almost identical to the unweighted sample. Thus, the ARMA(2, 1) and ARMA(1, 1) models are still the best candidate specifications.

Table 7 gives the weighted estimates for ARMA(1, 1) and ARMA(2, 1) specifications. The estimation is done using the same method-of-moments routine previously described. The only methodological difference between the two estimates (besides different sample sizes) is that the previous estimation programmed the statistical package to use a constant as the only instrument. The weighted estimation is done by designating in the statistical package the 1990 weights as instruments (which is the same as using a constant as an instrument but multiplying it by the weight).

The same balanced data are used here as were used in Sections 7.2 and 7.4, but this is somewhat misleading. The PSID assigns zero weights to many people in the sample for certain years, so the *effective* sample size for weighted regression falls from 959 to 720. Thus, there should be a slight widening of standard errors attributable to this sample size effect. Indeed, the weighted sample does have larger standard errors than the unweighted sample.

[46] Normally, one weights the entire moment condition. But when the only regressor is a constant the desired result can also be achieved by simply weighting the dependent variable. This convenient result turns on the fact that the weights are normalized to have an average value of one, thus they drop out of terms where they are multiplied by a constant.

Table 6
Covariogram based on weighted wage growth residuals

Statistics	Lag (order)						
	0	1	2	3	4	5	6
Constrained autocovariances (standard errors in parentheses)	0.112 (0.007)	−0.042 (0.004)	−0.004 (0.002)	−0.005 (0.003)	−0.002 (0.003)	−0.001 (0.003)	0.002 (0.003)
Autocorrelation	1	−0.37	−0.04	−0.04	−0.02	−0.01	0.02
Range of autocovariances	(0.101, 0.159)	(−0.061, −0.028)	(−0.011, 0.003)	(−0.006, 0.001)	(−0.014, 0.005)	(−0.007, 0.010)	(−0.007, 0.005)
Test for constant autocovariance (probability of event in parentheses)	yes (0.15)	yes (0.22)	yes (0.90)	yes (0.52)	yes (0.10)	yes (0.70)	yes (0.73)

Note. Based on seemingly unrelated regression model.

Table 7
Weighted estimates of ARMA processes for wage growth error structure

ARMA (p, q)	Estimation method	Direct estimates				Implied values from delta method	
		a_1	a_2	σ_{11}	σ_{12}	σ	m_1
(1, 1)	Estimating only autoregressive coefficients	−0.110 (0.0401)					
	Estimating autoregressive and covariance parameters jointly	−0.102 (0.0395)		0.119 (0.0079)	−0.052 (0.0042)	0.295 (0.0090)	−0.598 (0.0312)
(2, 1)	Estimating only autoregressive coefficients	−0.181 (0.0558)	−0.027 (0.0302)				
	Estimating autoregressive and covariance parameters jointly	−0.062 (0.0507)	0.025 (0.0260)	0.110 (0.0090)	−0.048 (0.0056)	0.286 (0.0090)	−0.587 (0.0569)

Note. Specifications based on Equations (7.6) and (7.8) (standard errors in parentheses).

Looking first at the results for the ARMA(1, 1) model, the a_1 parameter estimate is reasonably close to its unweighted value. Estimated alone or with the moving-average parameters it is about −0.11. In the joint estimation based on system (7.8), the covariance parameters are slightly lower in magnitude than in the unweighted sample and this shows up in a moving-average parameter of −0.598. This is only slightly smaller in magnitude than the estimates in the unweighted sample.

The most noteworthy difference from weighting is the effect that it has on the ARMA(2, 1) model autoregressive coefficient estimates. While estimation of just the autoregressive parameters (system (7.6)) gives similar results to those in Table 2, the coefficients change significantly when estimation includes the covariance parameters. The largest change shows up in the a_1 coefficient that moves from −0.186 to −0.062. This is beyond what one would expect and resembles the jump encountered in the bootstrap estimation, where the a_1 parameter in the full estimation also proves to be sensitive. Other parameter values do not change as much; the a_2 coefficient is of opposite sign but remains insignificant. The variance and autocovariance terms are both slightly lower in magnitude than earlier estimates. This leads to lower values of σ and m_1, 0.29 and −0.59. The unweighted estimates are not consistent for the same values as the weighted population and so there is no reason to expect them to be the same, but the joint estimation does appear to be sensitive to minor changes.

7.8. Results based on median regressions

The above empirical exercises investigate trends in mean wages, which involve the use of regression analysis. Such estimation techniques, of course, suffer from the fact that individual observations have unbounded influences upon the regression. Coding errors, reporting errors, and other anomalous events can have large effects on the estimated

coefficients. Such outliers might induce the sensitivity of the autoregressive parameters uncovered in the bootstrap and weighted estimators. Consequently, one might wish to consider bounded influence methods, such as median regression.[47]

Section 5 offers a convenient approach for estimating a smoothed version of median regression using standard nonlinear equation methods discussed in Sections 3 and 4. This subsection illustrates the use of these quantile regression procedures to estimate the parameters of the ARMA process governing the error structure of equations like (5.1). The models estimated below modify the specifications considered in Section 7.3 to estimate the error structure of (7.1). As such, although the computational problems of quantile regressions may involve a somewhat heavier burden due to their nonlinearities, the specifications are still relatively simple to implement.

An important difference between the estimation procedures outlined in Section 7.3 and those outlined below is that the researcher cannot use residuals $\widehat{U}_t$ instead of ω_t to simplify the estimation. Equation (5.1) specifies an equation where ω_t $(= y_t)$ depends on its own past values and a set of exogenous X_t's. Using mean regressions, as in Section 7.3, allows the researcher to estimate the error structure using the estimated residuals from regressing ω_t on the X_t's, due to the applicability of Proposition 4.1. This simplification cannot be used here – the smoothed median regression wraps the parameters into a cdf function Φ, and this nonlinearity leads to a violation of Proposition 4.1. Thus, median regressions require joint estimation of β and the a's.

The theory in Section 5 does not consider moving-average processes. Although it is possible to estimate a median version of Equation (7.8) when such a process is present – see Section 7.8.3 below – it is not entirely clear how the additional moving-average parameters are to be interpreted. Estimates are no longer consistent for autocovariances, but are instead consistent for some 'median' version of the autocovariances. Additionally, the easiest median specification requires assuming that, after accounting for the autoregressive component, the errors are uncorrelated across time. To illustrate this simple technique, this subsection starts with a purely autoregressive process even though the evidence presented above supports a moving average component. Given this focus, the estimation allows for the existence of a third autoregressive lag (ARMA(3, 0)), instead of just the first- or second-order models already considered.

7.8.1. Single equation estimation of a strictly autoregressive model

Section 5.2.1 outlines two approaches for estimating quantiles with panel data. The first and simplest assumes that error terms are independently distributed after one accounts for the autoregressive process. The second, using a system of equations, does not require this assumption for efficiency or correct inference.

For the sake of illustrating the first approach, assume the data satisfy an ARMA(3, 0) type process. Thus, one can consistently estimate coefficients using a structural equation

[47] As discussed in Section 5.2.2, one might also wish to use this type of estimator at several percentile points to better characterize the entire wage distribution [see Buchinsky (1994) for an example].

analogous to (5.8), which implies the following form for g_t:

$$f_t^{(16)} = g_t = \Phi\left(\frac{\omega_t - a_1\omega_{t-1} - a_2\omega_{t-2} - a_3\omega_{t-3} - X_t\beta}{S_N}\right) - (1 - \alpha_k), \qquad (7.9)$$

where the ω_t's and the X_t's are defined in Section 7.1. To estimate the median, set α_k equal to 0.5, and the smoothing parameter S_N equal to 0.03.[48] Φ is the standard normal cdf. The sample used to estimate Equation (7.9) consists of multiple years stacked as multiple observations. Thus, the sample size equals the number of people multiplied by the number of years in the panel. Setting $f_{ti} = g_{ti}$ in Equation (3.7) and using functions of the X_{ti}'s and all past values of ω_{ti} as the instrumental variables in formulating the Q_{ti}'s in (3.8), one can apply nonlinear two-stage least squares (N2SLS) to estimate the structural and autoregressive parameters of Equation (7.9).[49] If one suspects the structural errors g_{ti} are heteroscedastic, then one can select a robust option when computing standard errors.

As an alternative for increasing estimation efficiency, Sections 3.1.4, 3.2.2 and 4.1.4 outline the development of optimal instrumental variables. In the case considered here, due to the use of the normal cdf, the optimal instrument for a_1 takes the form

$$E\left(\left.\frac{\partial g_t}{\partial a_t}\right|_{\tilde{a}_1} | X_i, \omega_{t-1}, \ldots, \omega_{t-k}\right)$$
$$= E\left(-\frac{\omega_{t-1}}{S_N} * \phi\left(\frac{\omega_t - \tilde{a}_1\omega_{t-1} - \tilde{a}_2\omega_{t-2} - \tilde{a}_3\omega_{t-3} - X_t\tilde{\beta}}{S_N}\right)\right), \qquad (7.10)$$

where ϕ is the standard normal density. Designating $\gamma = (a_1, a_2, a_3, \beta)$, optimal instruments for each element of γ look similar; they are the standard normal density multiplied by the appropriate regressor. For (7.10) to constitute optimal instrumental variables in estimation requires structural errors, defined by g_{ti} in (7.9), to exhibit homoscedasticity across observations t and i. Given the nonlinear form of g_{ti} and its direct functional dependence on X_{ti}, satisfaction of this homoscedasticity assumption may be dubious.

Unfortunately, the expectations in (7.10) are unobserved and must be estimated. Given a consistent set of parameter estimates $\tilde{\gamma} = (\tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{\beta})$, form the estimated analog of (7.10), $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})$, the gradient vector of g_t, for each observation. To estimate an approximation for (7.10), regress $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})$ on flexible functions of the X_t's and all

[48] Setting S_N to other small values does not change the findings in any substantive way.

[49] Note that all lagged values of ω_t are predetermined since the model assumes there is no serial correlation beyond the autoregressive process already accounted for. Although one can use all past values of ω_t, a subset of early lags would also provide consistent estimates and may be more manageable to program and estimate. The important thing is to use instruments that provide the best predictive power while being uncorrelated with g_t.

past values of ω_t.[50] Call the fitted value of this regression $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})^P$; this quantity corresponds to the predicted value or projection of $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})$ conditioning on all exogenous and predetermined variables.

Using the $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})^P$ in place of the X_{ti}'s and past values of ω_{ti} as the instrumental variables in the application of N2SLS offers an alternative approach for estimating coefficients of Equation (7.9). This method improves efficiency in estimation assuming the g_{ti}'s are homoscedastic across observations. As recognized in 2SLS theory, if one were to use a series of flexible functional forms in the X_{ti}'s and past values of ω_{ti}, then one would expect little gain in efficiency in using the quantities $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})^P$ even with homoscedasticity since these functions would effectively span the space of these projected gradients.

The first row of Table 8 reports the results of estimating equation (7.9) using the optimal instrument set. The three autoregressive parameters are all large and significantly different from zero with point estimates of -0.397, -0.225, and -0.112. Note that these estimates are not necessarily consistent for the same coefficients estimated using mean regression, as these parameters are consistent for medians, not means. Regressed around the median, the autoregressive lag is much longer than its mean counterpart. The standard errors are all quite small, less than 0.02. So the smoothed median regression in this case appears to give precise results. Of course, these standard errors are constructed under the classical hypothesis of serially uncorrelated, homoscedastic error terms.

To check the validity of the asymptotic standard errors, Table 8 reports bootstrapped standard errors as well. The bootstrap, done in a manner similar to that in Section 7.5, yields identical standard errors to the asymptotically approximated errors. Note that this

Table 8
Quantile regression estimates of AR processes for wage growth error structure

ARMA (p, q)	Estimation method	a_1	a_2	a_3
(3, 0)	Individual regression	−0.397	−0.225	−0.112
	(asymptotic standard errors)	(0.019)	(0.017)	(0.014)
	(1000 reps bootstrap standard errors)	(0.019)	(0.017)	(0.016)
	System of equations	−0.463	−0.250	−0.065
	(asymptotic standard errors)	(0.010)	(0.007)	(0.006)
	(1000 reps bootstrap standard errors)	(0.027)	(0.022)	(0.024)
(3, 1)	System of equations	−0.637	−0.308	−0.131
	(asymptotic standard errors)	(0.014)	(0.010)	(0.009)
	(1000 reps bootstrap standard errors)	(0.069)	(0.028)	(0.025)

Note. Standard errors in parentheses.

[50] Due to the nonlinearity in g_t, the researcher might get better results by adding interactions and higher powers of the X_t's and lagged ω_t's.

bootstrap maintained the assumption of independence over time and so sampling was done on person-years, not on individuals; these bootstrapped standard errors and the estimates are inconsistent if there is serial correlation after accounting for the autoregressive process.

7.8.2. System of equations estimation of a strictly autoregressive process

While maintaining the assumption of an ARMA(3, 0), one can conceptually improve the efficiency of estimation by allowing for heteroscedasticity in structural disturbances across years for individuals. Moreover, one can allow for the possibility that coefficients are nonconstant over time. Instead of treating each person-year as an observation in Equation (7.9), one stacks these equations into a multiple equation system and treats each person as an observation of this system. This approach amounts to formulating a variant of f_{ti} given by (5.11), relying on Proposition 4.2 to compute estimates of the structural and autoregressive parameters of Equation (7.9).

When using functions of the X_{ti}'s and all past values of ω_{ti} as the instrumental variables, the procedure specifies the components of the f_{ti}'s as

$$\begin{aligned} f_t^{(17)} &= g_t, \\ f_t^{(18)} &= \hat{\eta}_{t-k} \times g_t, \quad k = 1, \ldots, (t - 1984),\ t = 1984, \ldots, 1990, \end{aligned} \tag{7.11}$$

where g_t is defined in Equation (7.9) with $\hat{\eta}_{t-k}$ calculated using Equation (5.10). Stacking these equations across years to obtain $f_{ti}(a_1, a_2, a_3, \beta) = (f_{ti}^{(17)\prime}, f_{ti}^{(18)\prime})'$ produces a model in the form of Equations (4.9) and (3.7). Estimation specifies all the elements of the X_{ti}'s as instrumental variables comprising Q_i appearing in (3.8).

One can again in principle achieve improved efficiency by exploiting an optimal set of instrumental variables given by (7.10). Approximating these quantities by the projections $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})^P$ discussed above, this use of instruments implies that one calculates $\hat{\eta}_{t-k}$ appearing in (7.11) using Equation (5.10) with $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})^P$ replacing the dependent variables y_{ti} and with Q_i, including a flexible set of quantities involving the X_{ti}'s providing for accurate approximations of the expected values of the gradients. Note that the vector $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})^P$ varies for each year, and is predicted with an ever-expanding set of predetermined values of ω_t; therefore the researcher cannot simply designate all the instruments as applying to all of the years. For example, $(\frac{\partial g_{1987}}{\partial \gamma}|_{\tilde{\gamma}})^P$ is predicted using ω_{1985}, which is correlated with g_{1985}. So $(\frac{\partial g_{1987}}{\partial \gamma}|_{\tilde{\gamma}})^P$ cannot serve as an instrument for the g_{1985} equation. If one were to introduce a full set of optimal instruments, then there will be as many $\hat{\eta}_{t-k}$ terms/equations in (7.11) for any given t as there exist parameters in the dynamic median equation (7.9). Attaining improved efficiency requires application of transformation (3.20) and the assumption of homoscedasticity across persons. Admittedly, the validity of this assumption may be suspect in light of the nonlinear character of the structural equation (7.9).

If one is not intending to attain full efficiency, then one can include only a subset of the $\hat{\eta}_{t-k}$ terms/equations in (7.11) associated with selected elements of $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})^P$. To

avoid having to include predetermined variables among the instrumental variable list used in estimation, a natural selection would be to include the terms corresponding to the elements linked to the autoregressive coefficients a_1, a_2 and a_3. The remaining parameters are identified through the incorporation of all the X_{ti}'s among the instrumental variable Q_i.

Depending on the assumptions maintained concerning the homoscedasticity of structural errors across individuals, nonlinear three-stage least squares (N3SLS) or method of moments (MM) offer procedures for estimating the structural and autoregressive coefficients of the dynamic median given by Equation (7.9). If one believes that errors are heteroscedastic, then use of the optimal weighting matrix in MM will increase efficiency over the single equation system estimated in Section 7.8.1. In the empirical illustration considered here, estimation relies on N3SLS and the variance–covariance matrix of the system of equations is assumed to satisfy homoscedasticity across individuals.

Row 2 of Table 8 reports the results from the application of N3SLS estimation of equation system (7.11) with $\hat{\eta}_{t-k}$ terms/equations incorporated corresponding to the a_1, a_2 and a_3 components of $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})^P$. Instrumental variables include all the exogenous variables X_i. Whereas the estimated value of a_2 in this row is similar to the findings obtained in the N2SLS estimation based on the single equation case discussed above, the estimate for a_1, at -0.463, is much higher in magnitude, especially given the tight standard error reported of 0.005. The a_3 coefficient drops to -0.065.

To examine the robustness of the standard errors, Table 8 also reports bootstrap standard errors. This bootstrap is done by sampling over the 959 individuals in the sample to preserve serial correlation. All three bootstrap standard errors are three times larger than their asymptotic counterparts, moving from around 0.007 to 0.024, implying that the asymptotic standard errors are misleading. Theoretically, (7.11) should be consistent for the same values as the estimation in Section 7.8.1, and may be more efficient. This requires the researcher to consider two questions. Why do these estimates from the system of equations not appear to be converging to the same value as those estimated in Section 7.8.1? And why are the standard errors reported on the supposedly more efficient procedure far larger?

The marked change in the parameter estimates may indicate misspecification. Specifically, if there is correlation across time in the g_t terms, then the first lagged residual, ω_{t-1}, is likely to be correlated with g_t, leading to inconsistent estimates. In this case the researcher should consider another specification – such as an ARMA(3, 1).

Although the bootstrap standard error estimates are larger for the system of equations than for the single-equation estimation, this is expected because the standard error estimates in Section 7.8.1 assume classical errors and are wrong in the presence of heteroscedastic errors or serial correlation. But the bootstrap standard errors reported in Section 7.8.1 are consistent under heteroscedasticity and they are identical to their asymptotic counterparts; thus the differences between the standard error estimates in Sections 7.8.1 and 7.8.2 are probably due to serial correlation. Section 7.8.2 imposes

neither homoscedasticity nor independence in computing standard errors.[51] Thus it is not surprising that the more efficient estimation reports larger standard errors, because the asymptotic standard errors calculated in Section 7.8.1 make much stronger assumptions. Given these findings, the researcher should consider another specification, such as an ARMA(3, 1), that is still consistent under the assumption of first-order serial correlation.

7.8.3. *Estimation of autoregressive coefficients allowing for a moving average component*

If the researcher wishes to allow for a first-order moving-average type process to be present in a quantile regression specification, then the system of equations introduced in Section 7.8.2 can be used to estimate autoregressive parameters with a slight modification. Under the assumption of an ARMA(3, 1), the approaches in Sections 7.8.1 and 7.8.2 are inconsistent because they treat the first lag in wage growth, ω_{t-1}, as a predetermined variable when in reality it is correlated with g_t.

Conceptually, a researcher could increase efficiency by using optimal instruments as in Section 7.8.2. In this case one would compute $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})^P$ from a least squares prediction of $(\frac{\partial g_t}{\partial \gamma}|_{\tilde{\gamma}})$ using a flexible set of regressors depending on X_i and all the predetermined values of ω_t (i.e., ω_{t-2} and those farther back). These predictions would then be used as described in Section 7.8.2.

Row 3 of Table 8 reports the results of N3SLS estimation of Equation (7.11) with instrumental variables incorporating the exogenous variables X_i's. The change in identifying assumptions does affect the estimates. The a_1 and a_2 parameter estimates are larger in magnitude than in the single equation procedure of Section 7.8.1: -0.637 vs. -0.397 and -0.308 vs. -0.225. The a_3 parameter, -0.131, is also larger in magnitude than it was in Section 7.8.1 or Section 7.8.2. These final estimates are consistent under heteroscedasticity and even under one-period serial correlation in the error terms.

Although the asymptotic standard errors imply minuscule confidence intervals, a bootstrap performed in the same manner as in the last subsection reveals standard errors that are, as before, three to five times larger than the asymptotically approximated standard errors. For example, the a_1 standard error estimate is 0.014 under the asymptotic approximation, while the bootstrap approximation is 0.069. Thus the asymptotic standard errors are not correctly approximating the small sample distribution. Note that these bootstrap standard errors are larger than those that resulted under the ARMA(3, 0) assumption. This is expected as ω_{t-1} is an important part of the identification strategy followed in Sections 7.8.1 and 7.8.2.

Although the median regression estimates do shift somewhat in response to assumptions about the specification, the qualitative implications are robust. The coefficients on

[51] Serial correlation would make the parameter estimates inconsistent due to the presence of correlated instruments, but the standard errors for these estimates would be correctly estimated.

the autoregressive parameters are about −0.40, −0.22, and −0.11. All three coefficients are statistically nonzero under either standard error estimate for all specifications. After adjusting for schooling and age, the average worker's wage growth is typically low in the years following a high growth year, and high in years following a low growth year. This is the same implication found using mean regression, but the autoregressive effect is stronger in medians than in means.

7.9. Summary of findings

The above empirical example characterizes the dynamic properties of hourly wages of men using the estimation procedures for time series models applied to panel data as described in Sections 3 and 4. Estimates of the covariogram associated with specification (7.1), and the autocorrelation coefficients implied by those estimates, accept the hypothesis of weak stationarity of the data. In addition, the sharp drop in the absolute value of the autocovariances after the first-order implies the existence of a first-order moving-average process. The gradual decline of the autocovariances of the second order and above also suggests a short autoregressive process. As a result, the mean regressions focused on estimating the autoregressive and moving average components of ARMA(1, 1) and ARMA(2, 1) time series processes in the errors.

Section 7.3 described procedures for estimating the autoregressive parameters alone and in conjunction with the moving-average parameters for the ARMA models implied by the results of Section 7.2. The pattern that emerges from Tables 2 and 3 is that the ARMA(1, 1) specification fits the data better than the ARMA(2, 1) specification does. Estimates of the first-order autoregressive parameter were similar across all versions of estimation, implying that 13 to 19 percent of residual wage growth dissipates the next year through the autoregressive parameter. The estimates for the second-order autoregressive parameter were small with standard errors of the same magnitude, leading to the conclusion that the ARMA(1, 1) model fits the data better than the ARMA(2, 1) does. Estimating the autoregressive parameters jointly with the moving-average parameters did not have any significant effect on the estimates of the autoregressive parameters and the moving-average parameters were tightly estimated. The first order moving-average parameter had a coefficient of about −0.65, implying that residual wage growth in one year typically reverts about halfway back in the next year.

Reliance on classical asymptotic standard errors for sample sizes as small as 1000 can potentially lead to overconfidence in one's inferences. The bootstrap standard errors calculated in Section 7.5 show that this is the case for the ARMA models estimated in Section 7.4. Specifically, bootstrap standard errors were larger than the classical asymptotic standard errors in all cases, and typically were about 25 percent larger. The second-order autoregressive parameter remained statistically indistinguishable from zero, and despite these changes the first-order autoregressive parameter remained statistically different from zero in all but one case. The bootstrap estimate of the first-order autoregressive parameter from the joint estimation of the ARMA(2, 1) procedure is significantly smaller than that given in Table 3. Viewed in light of the results of Section 7.7, it appears as

though joint estimation of the ARMA(2, 1) process is sensitive to changes in the procedure used.

The use of a balanced set of data in a panel context often requires that the researcher throw out individuals who do not have observations for every year of the sample. Section 7.6 re-estimates the covariogram and the autocorrelation coefficients associated with it using an unbalanced version of the data. Although comparisons of Table 5 with Table 1 show some difference in the estimates of the autocovariance parameters, the estimates presented in Table 5 still lead to the same conclusions: acceptance of the hypothesis of weak stationarity and the existence of a first-order moving-average process accompanied by a short autoregressive process in the errors.

Section 7.7 repeats the estimation reported in Sections 7.2 and 7.4 while accounting for the stratified sampling of the data through weighting. The relatively few quantitative or qualitative differences in these estimates should not be taken as an indication that weighting for a stratified sample is unimportant. For the particular PSID sample used here, Table 6 shows that weighting for the stratified nature of the data changes the covariogram quantitatively, but none of the qualitative conclusions change; the hypothesis of weak stationarity is maintained, as well as the existence of a first-order moving-average process and a short first- or second-order autoregressive process. Using the stratified sample weights to estimate the autoregressive parameters alone led to little change quantitatively or qualitatively versus the unweighted estimates. The weighted first-order autoregressive parameter estimates from the ARMA(1, 1) and ARMA(2, 1) models imply that 11 to 18 percent of residual variation in wage growth is undone the next year through the autoregressive parameter, as opposed to 13 to 19 percent for the unweighted counterpart. The weighted second-order parameter estimate for the ARMA(2, 1) model remains small and insignificant.

Joint estimation of the autoregressive and moving-average parameters using stratified sampling weights does lead to some important changes. Although the joint estimates of the ARMA(1, 1) coefficients using weights are similar to those from the unweighted procedure, the joint estimates for the ARMA(2, 1) model are different. Specifically, the estimate of the first-order autoregressive parameter is 1/3 the size of any of the other ARMA(2, 1) specifications, with a standard error that implies that the estimate is statistically indistinguishable from zero. In conjunction with the results of the bootstrap estimates of Section 7.5, it appears that joint estimation of the ARMA(2, 1) process is sensitive to changes in the procedure used.

Turning to the median regressions, one can limit the sensitivity of the estimation to outliers in a tractable way by using the techniques developed in Section 5. This smoothed median regression estimator can be done with one equation or several, and can be thought of as a fairly standard nonlinear optimization problem. The estimation revealed a three period lag in the autoregressive parameters with coefficients of -0.397, -0.225, and -0.112, respectively. Single and multiple equation techniques are illustrated, both of which can be estimated using standard techniques of two and three stage least squares or method of moments. The single equation estimator reports identical standard errors under asymptotic and bootstrap approximations. The same cannot be

said for the system of equations methods that use an optimal weighting matrix. Judging from the bootstrapped standard errors, the asymptotic standard errors reported here under report the true variance in the sampling distribution by a factor of three.

The techniques executed above show how to apply fairly sophisticated methods of estimation within a very simple estimation framework. Any software package capable of estimating a nonlinear system of equations (preferably with instrumental variables, although this is not required) should be able to perform all the techniques introduced here. The standard errors reported by these routines are asymptotically valid.

Bootstrapped standard errors indicate that the asymptotic estimates may be a useful guide, but could deviate substantially from the true values. This is especially true in a median regression of systems of equations where one uses an optimal weighting matrix. Thus, there is a potential gain from bootstrapping estimates in systems of equations.

8. Summary and concluding remarks

The goal of this chapter has been to present a unified set of estimation methods for fitting a rich array of models describing dynamic relationships within a longitudinal data setting. The chapter is motivated by the principle that, whenever possible, these methods should rely on routines available in familiar software packages to make them accessible to a broad range of practitioners. The discussion covers both the empirical specifications and estimation methods applicable in a wide variety of longitudinal analyses. The exposition motivates approaches by considering applications aimed at characterizing the intertemporal properties of wages and earnings, a research area in which one finds virtually all assortments of longitudinal applications. In addition to presenting the econometric principles underlying approaches, this chapter illustrates methods through a series of empirical examples using hourly wages data on men from the PSID.

As outlined in Section 2, panel data specifications designed to capture the underlying sources of micro dynamics experienced by individuals consist of two components: (i) parametric relationships depicting the links among current, past and future values of measured variables, be they endogenous, predetermined or exogenous quantities; and (ii) error structures describing the stochastic properties of disturbances introduced in relationships to account for unmeasured factors. Nonlinear simultaneous equation specifications provide a general class of models for the relationships linking measured variables. Especially useful simplifications include dynamic simultaneous equations models (DSEM) incorporating rational distributed lags that allow researchers to entertain flexible lag structures having finite or infinite order using short time series of the sort available in longitudinal data.

Popular formulations for error structures include variants of autoregressive-moving average (ARMA) processes. In a panel data setting, a researcher enjoys a wider choice of specifications because distributed lag and ARMA parameters can be permitted to vary freely over time. Furthermore, error specifications provided by ARMA schemes can be readily extended to incorporate permanent and random trend error components. Special

problems arise in deriving parameterizations of the variance–covariance matrix associated with ARMA processes in a longitudinal data setting. These problems pertain to the treatment of initial conditions, which are particularly troublesome for mixed ARMA specifications. Section 4.3 proposes a general solution to this problem. The DSEM and these extensions encompass most of the specifications found in the longitudinal literature.

The "method of moments" (MM) framework conceptually provides a general approach for estimating parameters of panel data specifications, and Section 3 outlines the particular formulations and key asymptotic results relied upon in this framework for computing estimates and testing hypotheses. The discussion summarizes approaches in the literature for exploiting predetermined variables as instrumental variables in the MM framework, as well as selecting instrumental variables that yield the greatest efficiency.

Section 4 covers several specializations of the MM approach that can substantially simplify the problem of estimating sophisticated specifications or many equations in a longitudinal data context. One application includes linear/nonlinear 3SLS procedures, a well-known special case of the MM framework that yields convenient computational formulas for large systems of equations. While conventional implementation of 3SLS routines do not permit use of predetermined variables as instruments, Section 4.1 demonstrates how one can readily overcome this shortcoming by adding new structural equations to the model while staying within a standard 3SLS program.

There are also considerable advantages to breaking up a longitudinal data estimation problem into parts, allowing researchers to focus on one part of the model at a time. The panel data models introduced in Section 2 provide a rich set of specifications, making the task of choosing among these specifications a formidable endeavor. Not only do they permit flexible parameterizations relating measured variables, but numerous formulations are available for error processes; indeed, far more than can be entertained in standard time series analyses. A researcher rarely knows precisely which parameterizations are consistent with the data, and typically must invest considerable effort in performing diagnostic procedures designed to narrow model choices.

In view of this complexity, Section 4 presents a variety of procedures allowing researchers to subdivide the problem of estimating parameters of sophisticated longitudinal specifications into a multi-stage approach. One can estimate parameters determining the autocovariance patterns of errors separately from the structural coefficients directly associated with measured variables, as well as further separating estimation of parameters of the AR and the MA components of the error process. In each step, the application of familiar estimation routines reports valid test statistics that are useful for discovering which parts of a model fit the data without having to specify all parts together. Moreover, these procedures offer a powerful set of diagnostic tools useful not only for evaluating the basic features of specifications – such as identifying the orders of ARMA models consistent with the data – but also for discovering reliable values for parameters that can serve as starting values for the larger estimation exercises.

Section 5 considers using conditional quantile regressions to describe the dynamics of earnings, a set of empirical specifications representing an attractive alternative

to DSEMs. The analysis considers the formulation of quantiles analogous to autoregressive models, including systems of equations permitting one to evaluate how several different percentiles jointly evolve over time. The section further describes a flexible approach for estimating the coefficients of autoregressive quantile equations by implementing conventional nonlinear instrumental variable procedures. Thus, the estimation approaches and issues considered throughout this chapter apply to computing estimates and test statistics for these dynamic quantile specifications as well.

Section 6 describes how to incorporate weighting and unbalanced data in the estimation of longitudinal data models, which is applicable for both linear and nonlinear specifications. The type of weighting considered in this discussion corresponded to the sort typically provided in survey data to account for stratified sampling designs implemented during data collection, designs which produce nonrandom samples. Not only must the construction of weights account for the stratification of the original sample, this construction must also adjust for the sample attrition which contributes to a varying sample composition over time. Virtually all survey data sources contain such weights, and not using them in estimation produces inconsistent estimates of even basic statistics. Naive use of weighting options available in standard software packages also generates incorrect calculations for standard errors and test statistics. The discussion documents how one must modify the MM formula to account for stratified sampling. The section ends by describing a modified weighting-type procedure enabling one to use conventional methods to estimate intertemporal specifications with unbalanced data, which are samples supplying an imperfect overlap in the time periods available for individuals included in the longitudinal survey. The procedures covered in this discussion also apply to estimating the parameters of conditional quantile relationships using stratified and/or unbalanced samples.

To illustrate the estimation approaches covered in this chapter, Section 7 applies many of the methods in an empirical analysis of the dynamic properties of the hourly wages of men during the period 1980–1991 using data from the PSID. While this analysis merely provides examples of methods to highlight critical concepts, comparisons of findings across procedures offers insights into how various procedures influence results. Estimates of the covariogram using data on residuals support the hypothesis of weak stationarity for wage growth, with the pattern of estimated autocovariances and test statistics suggesting that an ARMA(1, 1) model adequately describes the data. Applying procedures that estimate parameters of the AR and MA portions of this model in separate steps yields values for the coefficients similar to those obtained by joint estimation of the parameters of the ARMA(1, 1) specification.

The empirical analysis goes on to examine the sensitivity of results to using: (1) bootstraps to calculate estimates and standard errors, (2) unbalanced data, and (3) weighting to account for stratified sampling inherent in survey data. According to the findings, bootstrap standard errors tend to exceed those based on classical asymptotic theory, typically being about 25 percent larger. The calculation of estimates using unbalanced data allows a researcher to exploit all data available for a person, without requiring deletion of individuals who do not have observations for every year of the sample. Although

results for the balanced and unbalanced data show differences, both sets of findings still allow acceptance of the hypothesis of weak stationarity and the underlying presence of an ARMA(2, 1) process. Finally, joint estimation of the autoregressive and moving-average parameters using stratified sampling weights does lead to substantial changes for some estimates of the ARMA(2, 1) model.

Section 7 also illustrates the estimation of dynamic quantiles, focusing on the intertemporal variation in medians. The empirical analysis reveals the presence of at least a three-period lag in the autoregressive structure in the median of wages. The application of bootstrap procedures yields substantially larger standard errors for the multi-equation estimation methods, but not for the single equation approach.

Other chapters in the *Handbook of Econometrics* and the *Handbook of Labor Economics* offer valuable alternative or complementary discussions of the topics covered here. In the area of estimation approaches applicable for panel data, Chamberlain (1984) has become a standard reference, and Arellano and Honoré (2001) provides a thoughtful update of recent developments. Horowitz (2001) discusses the theoretical underpinnings for the bootstrap procedures pertinent to the estimation methods described in this study. Variants of the empirical models discussed in this chapter also appear in the body of work surveyed by Solon (1999), which summarizes what has been learned from recent research on intergenerational earnings mobility. Beyond the Handbooks, the textbooks by Hsiao (1986, 2003) and Baltagi (1995) provide comprehensive reviews of the panel data literature, offering a wealth of references and detailed presentations of many concepts only touched on in this chapter.

Econometric developments introduced to analyze longitudinal data comprise one of the most active research areas in the past three decades. No doubt these developments will continue since these data constitute the richest sources of information available to economists hoping to understand a wide range of phenomena. Just as in the past, the study of wage, earnings and income dynamics will motivate many of these econometric innovations.

Appendix A: Specifying the covariance matrix for an ARMA process

The purpose of this appendix is to provide explicit parameterizations for the covariance matrix $E\{U_i U_i'\}$ associated with the vector of transitory components. The assumptions and notation introduced to derive the specification for $E(U_i U_i'\}$ given by relations (4.16)–(4.19) are also used here. The following discussion begins with the development of a simple parameterization corresponding to relations (4.16)–(4.19).

Using (4.14), it is possible to reformulate the system of equations given by (4.13). To avoid the need for dealing with several possible cases, it is convenient to introduce the notation $\zeta_j = 0$ for $j < 0$ (for $j = 0$, $\zeta_0 \equiv 1$ and for $j > 0$, $\zeta_j \equiv m_j - \sum_{h=1}^{j} a_h \zeta_{j-h}$) and the definition that a summation of the form $\sum_{h=0}^{c}$ is equal to zero whenever $c < 0$.

Using this notation, Equations (4.13) and (4.14) imply

$$\begin{bmatrix} \sum_{j=0}^{p} a_j U_{(T-j)i} \\ \vdots \\ \sum_{j=0}^{p} a_j U_{(p+1-j)i} \\ U_{pi} \\ U_{(p-1)i} \\ \vdots \\ U_{1i} \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^{q} m_j \varepsilon_{(T-j)i} \\ \vdots \\ \sum_{j=0}^{q} m_j \varepsilon_{(p+1-j)i} \\ \sum_{j=0}^{q-1} \zeta_j \varepsilon_{(p-j)i} \\ \sum_{j=0}^{q-1} \zeta_{j-1} \varepsilon_{(p-j)i} \\ \vdots \\ \sum_{j=0}^{q-1} \zeta_{j-p+1} \varepsilon_{(p-j)i} \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mu_{pi} \\ \mu_{(p-1)i} \\ \vdots \\ \mu_{1i} \end{bmatrix}, \tag{A.1}$$

where

$$\mu_{ti} = \sum_{j=t-p+q}^{t-\ell_1} \zeta_j \varepsilon_{(t-j)i} + \sum_{j=t-\ell_1+1}^{t-\ell_2} \zeta_j \phi_{(t-j)i}, \quad t = 1, \ldots, p.$$

The first set of $T - p$ equations in (A.1) is simply the standard representation of the ARMA process generating $U_{(p+1)i}, \ldots, U_{Ti}$, and the second set of p equations is the moving average representation of the ARMA process for $U_{1i}, \ldots, U_{pi}$ with the μ_{ti}'s, $t = 1, \ldots, p$, defined to include all disturbances realized prior to period $p - q + 1$. The formulation of (A.1) assumes that $\ell_1 < p - q + 1$.[52] In matrix notation, (A.1) may be written as

$$FU_i = G \begin{pmatrix} \varepsilon_i \\ \mu_i \end{pmatrix}, \tag{A.2}$$

where

$$U_i = \begin{pmatrix} U_{Ti} \\ \vdots \\ U_{1i} \end{pmatrix}, \quad \varepsilon_i = \begin{pmatrix} \varepsilon_{Ti} \\ \vdots \\ \varepsilon_{(p-q+1)i} \end{pmatrix}, \quad \mu_i = \begin{pmatrix} \mu_{pi} \\ \vdots \\ \mu_{1i} \end{pmatrix}.$$

F is the $T \times T$ matrix defined below (4.15), and G is a $T \times (T + q)$ matrix defined as

$$G = \begin{bmatrix} M_{(T-p)\times(T-p+q)} & O_{(T-p)\times p} \\ O_{p\times(T-p)} \; K_{P\times q} & I_p \end{bmatrix}.$$

M is a diagonal band matrix with the elements $(m_0, \ldots, m_q)$ running down the diagonal,[53] and the matrix K has $(\zeta_0, \ldots, \zeta_{q-1})$ as its first row, $(0, \zeta_0, \ldots, \zeta_{q-2})$ as its second row, and so on until the pth row is reached, or if $q < p$, until the qth row is reached, after which the rows of K contain zeros. When forming the partitioned matrices associated with F and G, the above analysis assumes that any matrix with an implied

[52] The justification for the restriction can be found in the footnote following Equation (4.18).

[53] A diagonal band matrix is specified in the first footnote following Equation (4.15).

dimension equal to zero is deleted from the specification. Thus, when $p = 0$, $F = [A]$ and $G = [M]$; and when $q = 0$, K is eliminated and

$$G = \begin{bmatrix} M & O \\ O & I \end{bmatrix}.$$

Given the expression for U_i implied by (A.2), the problem of parameterizing $E\{U_i U_i'\}$ becomes one of specifying a correlation structure for disturbance vectors ε_i and μ_i. Since each of the components of ε_i follows a white noise error process, we have

$$E\{\varepsilon_i \varepsilon_i'\} = (I_{T-p+q} \otimes \sigma^2) \equiv \Sigma, \tag{A.3}$$

where $\sigma^2 = E\{\varepsilon_{ti}^2\}$ for $t = p - q + 1, \ldots, T$. Inspection of the formulas for the μ_{ki}'s reveals three facts: (i) the μ_{ki}'s depend on a common set of disturbances; (ii) all of these disturbances are realized prior to period $p - q + 1$; and (iii) included among these disturbances are the initial conditions for the ARMA process (i.e., the ϕ_{ki}'s). Since each of the components of ε_i are realized during and after period $p - q + 1$, fact (ii) implies $E(\mu_i \varepsilon_i') = 0$. Fact (i) implies that the components of μ_i are mutually correlated, so $E\{\mu_i \mu_i'\}$ contains no zero elements in general. In addition, without imposing rigorous conditions, fact (iii) indicates that no restrictions will exist on the form of $E\{\mu_i \mu_i'\}$. In general, then, μ_i will possess an arbitrary covariance structure which we may formally express as

$$E\{\mu_i \mu_i'\} = \Lambda, \tag{A.4}$$

where Λ is any positive definite, symmetric matrix.

Combining the above results, we obtain the following specification for $E\{U_i U_i'\}$:

$$\Theta = F^{-1} G \begin{bmatrix} \Sigma & O \\ O & \Lambda \end{bmatrix} G' F^{-1\prime}. \tag{A.5}$$

This parameterization imposes all of the restrictions implied by the ARMA process unless one is willing to introduce precise information about how and when this process started.

There are two modifications of the above parameterization that may be useful in applied work. First, to simplify the construction of the matrix K, one may replace each of the nonzero elements of this matrix (i.e., all the ζ_j's, $j \geqslant 0$) by arbitrary parameters, rather than using the coefficients of the ARMA process and the formulas specified above to form these elements. This modification avoids the need for imposing nonlinear restrictions, but it introduces new parameters and reduces the efficiency of estimation.

The second modification concerns the parameterization of Λ defined by (A.4). This matrix is purely a theoretical construct and represents nuisance parameters. An unattractive feature of this parameterization is that one cannot easily infer an approximate value for Λ using preliminary data analysis techniques or estimation methods that do not require the full estimation of Θ. An alternative specification is obtained if one replaces the matrix Λ by the matrix $\Upsilon = E\{\mu_i \mu_i'\} + K(I_q \times \sigma^2)K'$ which is also only restricted

to be positive definite and symmetric. Substituting this new parameterization into (A.5) implies

$$\Theta = F^{-1}\begin{bmatrix} M\Sigma M' & M\Sigma\begin{bmatrix} O' \\ K' \end{bmatrix} \\ [OK]\Sigma M' & \Upsilon \end{bmatrix} F^{-1\prime}. \tag{A.6}$$

According to this new specification, $\Upsilon = E\{U_{(1)i}U'_{(1)i}\}$, where the vector $U'_{(1)i} \equiv (U_{pi}, \ldots, U_{1i})$ includes the last p components of U_i. In contrast to the previous parameterization, Υ can be estimated prior to the full estimation of Θ. The specification given by (A.6) is equivalent to the one implied by relations (4.16)–(4.19) presented in Section 4.

It is straightforward to generalize the above specifications to deal with the case where more than one structural equation is included in model (4.10) (i.e., there are several equations for each period) and where the disturbance vector U_{ti} follows a multivariate ARMA process. One merely needs to replace the coefficients a_j, m_j, and ζ_j in the above specifications by matrices with dimensions equal to the number of equations and redefine the dimensions of other matrices so that they are conformable.

Appendix B: A general approach for estimating ARMA processes

This appendix presents a general framework for estimating parameters of a stationary multivariate autoregressive moving-average (ARMA) process applying the procedures summarized in Sections 4, 6 and 7.

Consider the following stationary multivariate ARMA process

$$F(L)V_t = \sum_{j=1}^{J} M_j(L)\varepsilon_{tj}, \tag{B.1}$$

where V_t, is a vector of observed variables with zero mean, $F(L) = F_0 + F_1L + \cdots + F_pL^p$ and $M_j(L) = M_{j0} + M_{j1}L + \cdots + M_{jq}L^q$ are matrix lag polynomials, and the ε_{tj} are mutually independent, white-noise error vectors with

$$E\left(\varepsilon_{tj}\varepsilon'_{tj}\right) = \Sigma_j, \quad j = 1, \ldots, J.$$

The inclusion of more than one moving average component in relation (B.1) allows the model to incorporate error-component specifications. Obviously, if error components are present, then the forms of the $F(L)$ and $M_j(L)$ lag matrices must be restricted according to some structural constraints to achieve identification of coefficients.

Represent the kth order autocovariances of V_t, by

$$\Theta_{-\kappa} = E\left(V_tV'_{t-k}\right).$$

The stationarity of the process implies that $\Theta_\kappa = \Theta_{-\kappa}$.

Using multivariate extensions of equations (9), (10), and (11) in Section 5.8 of Anderson (1971), one can derive from (B.1) the following system of equations:

$$\begin{aligned}
&\sum_{i=0}^{p}\left(\sum_{h=0}^{p} F_h \Theta_{h-i}\right) F_i' = \sum_{j=0}^{J}\sum_{k=0}^{q} M_{jk}\Sigma_j M_{jk}',\\
&\sum_{i=0}^{p}\left(\sum_{h=0}^{p} F_h \Theta_{h-i-1}\right) F_i' = \sum_{j=0}^{J}\sum_{k=0}^{q-1} M_{j,k+1}\Sigma_j M_{jk}',\\
&\quad\vdots\\
&\sum_{i=0}^{p}\left(\sum_{h=0}^{p} F_h \Theta_{h-i-q}\right) F_i' = \sum_{j=0}^{J} M_{jq}\Sigma_j M_{j0}',\\
&\sum_{h=0}^{p} F_h \Theta_{h-q-1} = 0,\\
&\quad\vdots\\
&\sum_{h=0}^{p} F_h \Theta_{h-T+1} = 0,
\end{aligned} \tag{B.2}$$

where T is the farthest time period available in the longitudinal data set for computing autocovariances (assumed to satisfy $T - 1 > q$). Let θ be a vector which stacks the own and cross-autocovariances in (B.2), and let α be a parameter vector in which the unknown elements of the matrices F_j, M_{jk} and Σ_j are stacked. With some algebraic manipulations, the system of equations in (B.2) can be stacked to yield a vector equality of the form

$$f(\alpha, \theta) = 0, \tag{B.3}$$

with evaluation of this relationship at the true values of the parameters.

To understand how $f(\alpha, \theta)$ is formed, first consider the case of a univariate ARMA process. Then

$$\begin{aligned}
f_1(\alpha, \theta) &= \sum_{i=0}^{p}\sum_{h=0}^{p} F_h F_i \Theta_{h-i} - \sum_{j=0}^{J}\sum_{k=0}^{q} M_{jk}^2 \sigma_j^2,\\
f_2(\alpha, \theta) &= \sum_{i=0}^{p}\sum_{h=0}^{p} F_h F_i \Theta_{h-i-1} - \sum_{j=0}^{J}\sum_{k=0}^{q-1} M_{j,k+1} M_{jk} \sigma_j^2,
\end{aligned}$$

and so on. In the case of an n-variate ARMA process, recognize that the matrix

$$\sum_{i=0}^{p}\left(\sum_{h=0}^{p} F_h \Theta_{h-i}\right) F_i' - \sum_{j=0}^{J}\sum_{k=0}^{q} M_{jk}\Sigma_j M_{jk}'$$

is symmetric. Hence, to form $f(\alpha, \theta)$ one stacks only the upper triangular part of this matrix. Thus, the first $(n(n+1)/2)$ elements of $f(\alpha, \theta)$ are generated by the first matrix equation in (B.2), the next n^2 are generated by the second matrix equation in (B.2), and so on.

Let α^* and θ^* denote the true values of α and θ. Further, suppose that $\hat{\theta}$ is a consistent estimate of θ^* with $\sqrt{N}(\hat{\theta} - \theta^*)$ converging in distribution to $N(0, H)$, where N is the sample size. Then, as long as $f(\alpha, \theta)$ satisfies the general conditions given in Section 3, it is the case that

$$\sqrt{N} f(\alpha^*, \hat{\theta}) \xrightarrow{d} N\left(0, \left(\left.\frac{\delta f}{\delta \theta'}\right|_{\alpha^*, \theta^*} H \left.\frac{\delta f'}{\delta \theta}\right|_{\alpha^*, \theta^*}\right)\right). \tag{B.4}$$

When $(\delta f / \delta \alpha')$ has full column rank, it will be possible in (B.3) to solve for the elements of α in terms of the elements of θ. Since θ will typically be overidentified, the result in (B.4) justifies the applicability of MM procedures to compute a unique estimate of α. This procedure calculates an estimate α to minimize the function

$$Q = f(\alpha, \hat{\theta})' W f(\alpha, \hat{\theta}), \tag{B.5}$$

where W is any positive definite matrix. The resulting estimate is consistent and asymptotically normal given (B.4). Further, if $\widehat{H}$ is a consistent estimate of H and $\hat{\alpha}$ is a consistent estimate of α, the result in (B.4) enables one to conclude that $[\frac{\delta f}{\delta \theta'}|_{\tilde{\alpha}, \hat{\theta}} \widehat{H} \frac{\delta f'}{\delta \theta}|_{\tilde{\alpha}, \hat{\theta}}]^{-1}$ is an optimal choice for W. Standard MM procedures then imply that

$$\hat{\alpha} \sim N\left(\alpha^*, \frac{1}{N}\left(\left.\frac{\delta f'}{\delta \alpha}\right|_{\hat{\alpha}, \hat{\theta}} \left(\left.\frac{\delta f}{\delta \theta'}\right|_{\hat{\alpha}, \hat{\theta}} \widehat{H} \left.\frac{\delta f'}{\delta \theta}\right|_{\hat{\alpha}, \hat{\theta}}\right)^{-1} \left.\frac{\delta f}{\delta \alpha'}\right|_{\hat{\alpha}, \hat{\theta}}\right)^{-1}\right), \tag{B.6}$$

where $\hat{\alpha}$ is the estimate of α when W is chosen optimally.

The value of the function Q given in (B.5), with W chosen optimally, forms the basis for a statistic to test whether the autocovariances of V_t, have a parameterization implied by (B.1). Let $\widehat{Q}$ be the value of the function Q, let k_θ denote the number of elements in θ, and let k_α denote the number of parameters contained in α. According to the findings in (3.17), it follows that if the null hypothesis given by (B.3) is true, then $N\widehat{Q}$ is approximately distributed as a chi-squared random variable with $(k_\theta - k_\alpha)$ degrees of freedom. This statistic provides a measure of fit of the parameterized multiple time series model to the sample own- and cross-covariograms, with the alternative hypothesis interpreting all variances and autocovariances as being entirely unconstrained.

References

Altonji, J., Dunn, T. (2000). "An intergenerational model of wages, hours, and earnings". Journal of Human Resources 35, 221–258.

Amemiya, T. (1974). "The nonlinear two-stage least-squares estimator". Journal of Econometrics 2, 105–110.

Amemiya, T. (1975). "The nonlinear limited-information maximum-likelihood estimator and the modified nonlinear two-stage least-squares estimator". Journal of Econometrics 3, 375–386.
Amemiya, T. (1977). "The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model". Econometrica 45, 955–968.
Amemiya, T. (1983). "Non-linear regression models". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. 1. North-Holland, Amsterdam. Chapter 6.
Amemiya, T. (1985). Advanced Econometrics. Harvard University Press, Cambridge.
Amemiya, T., MaCurdy, T.E. (1986). "Instrumental-variable estimation of an error-components model". Econometrica 54, 869–881.
Anderson, T.W. (1971). The Statistical Analysis of Time Series. John Wiley and Sons, New York.
Anderson, T.W., Hsiao, C. (1982). "Formulation and estimation of dynamic models using panel data". Journal of Econometrics 18, 47–82.
Arellano, M., Bond, S.R. (1991). "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations". Review of Economic Studies 58, 277–297.
Arellano, M., Bover, O. (1995). "Another look at the instrumental-variable estimation of error-components models". Journal of Econometrics 68, 29–51.
Arellano, M., Honoré, B. (2001). "Panel data models: Some recent developments". In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. V. North-Holland, Amsterdam. Chapter 53.
Ashenfelter, O. (1978). "Estimating the effects of training programs on earnings". Review of Economics and Statistics 40, 47–57.
Baker, M., Solon, G. (2003). "Earnings dynamics and inequality among Canadian men, 1976–1992: Evidence from longitudinal income tax records". Journal of Labor Economics 21, 289–321.
Baltagi, B.H. (1995). Econometric Analysis of Panel Data. John Wiley and Sons, Chichester.
Baltagi, B.H. (Ed.) (2002). Recent Developments in the Econometrics of Panel Data. American International Distribution Corporation, Williston.
Bhargava, A., Sargan, J.D. (1983). "Estimating dynamic random effects models from panel data covering short time periods". Econometrica 51, 1635–1659.
Buchinsky, M. (1994). "Changes in the US wage structure 1963–1987: Application of quantile regression". Econometrica 62, 405–458.
Chamberlain, G. (1982). "Multivariate regression models for panel data". Journal of Econometrics 18, 5–46.
Chamberlain, G. (1984). "Panel data". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. II. North-Holland, Amsterdam. Chapter 22.
David, M. (1971). "Lifetime income variability and income profiles". In: Proceedings of the Annual Meeting of the American Statistical Association, pp. 285–292.
Fitzgerald, J., Gottschalk, P., Moffitt, R. (1998). "An analysis of sample attrition in panel data: The Michigan panel study of income dynamics". NBER Working Paper No. t0220.
Friedman, M., Kuznets, S. (1945). Income from Independent Professional Practice. National Bureau of Economic Research, New York.
Geweke, J. (1984). "Inference and causality in economic time series models". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. II. North-Holland, Amsterdam. Chapter 19.
Geweke, J., Keane, M. (2001). "Computationally intensive methods for integration in econometrics". In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. V. North-Holland, Amsterdam. Chapter 56.
Granger, C.W.J., Watson, M.W. (1984). "Time series and spectral methods in econometrics". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. II. North-Holland, Amsterdam. Chapter 17.
Hajivassiliou, B.A., Ruud, P.A. (1994). "Classical estimation methods for LDV models using simulation". In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, Amsterdam. Chapter 40.
Hamilton, J.D. (1994). "State-space models". In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, Amsterdam. Chapter 50.
Hannan, E.J. (1969). "The identification of vector mixed autoregressive moving-average systems". Biometrika 56, 223–225.

Hansen, L.P. (1982). "Large sample properties of generalized method of moment estimators". Econometrica 50, 1029–1054.
Härdle, W., Linton, O. (1994). "Applied nonparametric methods". In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, Amsterdam. Chapter 38.
Hause, J. (1977). "The covariance structure of earnings and the on-the-job training hypothesis". Annals of Economic and Social Measurement 6, 73–108.
Hause, J. (1980). "The fine structure of earnings and the on-the-job-training hypothesis". Econometrica 48, 1013–1029.
Heckman, J.J., Singer, B. (1986). "Econometric analysis of longitudinal data". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. III. North-Holland, Amsterdam. Chapter 29.
Hendry, D.F., Pagan, A.R., Sargan, J.D. (1984). "Dynamic specification". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. II. North-Holland, Amsterdam. Chapter 18.
Hill, M. (1992). The Panel Study of Income Dynamics: A User's Guide. Sage Publications, Newbury Park.
Holtz-Eakin, D., Newey, W., Rosen, H. (1988). "Estimating vector autoregressions with panel data". Econometrica 56, 1371–1395.
Horowitz, J.L. (2001). "The bootstrap". In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. V. North-Holland, Amsterdam.
Hsiao, C. (1986). Analysis of Panel Data. Cambridge University Press, London.
Hsiao, C. (2003). Analysis of Panel Data, second ed. Cambridge University Press, London.
Kashyap, R.L., Nasburg, R.E. (1974). "Parameter estimation in multivariate stochastic difference equations". IEEE Transactions on Automatic Control 19, 784–797.
Keane, M., Runkle, D. (1992). "On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous". Journal of Business and Economic Statistics 10, 1–9.
Lillard, L., Weiss, Y. (1979). "Components of variation in panel earnings data: American scientists 1960–1970". Econometrica 47, 437–454.
Lillard, L., Willis, R. (1978). "Dynamic aspects of earnings mobility". Econometrica 46, 985–1012.
MaCurdy, T. (1982a). "Using information on the moments of disturbances to increase the efficiency of estimation". NBER Working Paper No. t0022.
MaCurdy, T., Hong, H. (1998). "Smoothed quantile regression in generalized method of moments". Stanford Working Paper.
Manski, C. (1994). "Analog estimation of econometric models". In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam. Chapter 43.
McFadden, D.L. (1984). "Econometric analysis of qualitative response models". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. II. North-Holland, Amsterdam. Chapter 24.
Powell, J.L. (1994). "Estimation of semiparametric models". In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, Amsterdam. Chapter 41.
Sargan, J.D. (1958). "The estimation of economic relationships using instrumental variables". Econometrica 26, 393–415.
Sargan, J.D. (1961). "The maximum likelihood estimation of economic relationships with autocorrelated residuals". Econometrica 29, 414–426.
Stock, J.H. (1994). "Unit roots, structural breaks and trends". In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, Amsterdam. Chapter 46.
Teräsvirta, T., Tjøstheim, D., Granger, C.W.J. (1994). "Aspects of modelling nonlinear time series". In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, Amsterdam. Chapter 48.
van den Berg, G.J. (2001). "Duration models: Specification, identification, and multiple durations". In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. V. North-Holland, Amsterdam. Chapter 55.
Watson, M.W. (1994). "Vector autoregression and cointegration". In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, Amsterdam. Chapter 47.
Wooldridge, J.M. (1994). "Estimation and inference for dependent processes". In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, Amsterdam. Chapter 45.

Further reading

Abadie, A. (2000). "Semiparametric estimation of instrumental variable models for causal effects". NBER Working Paper No. t0260.

Abadie, A., Angrist, J.D., Imbens, G.W. (1998). "Instrumental variables estimation of quantile treatment effects". NBER Working Paper No. t0229.

Abbring J., Heckman, J.J., Vytlacil E. (2007). "Econometric evaluation of social programs, Parts I–III". In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. North-Holland, Amsterdam. In press (Chapters 70–72).

Abowd, J., Card, D. (1984). "Intertemporal substitution in the presence of long term contracts". Working Paper No. 166. Industrial Relations Section, Princeton University.

Abowd, J., Card, D. (1989). "On the covariance structure of earnings and hours changes". Econometrica 57, 411–445.

Abowd, J., Crépon, B., Kramarz, F. (1997). "Moment estimation with attrition". NBER Working Paper No. t0214.

Ahn, S.C., Schmidt, P. (1995). "Efficient estimation of models for dynamic panel data". Journal of Econometrics 68, 5–27.

Ahn, S.C., Schmidt, P. (1999). "Modified generalized instrumental variables estimation of panel data models with strictly exogenous instrumental variables". In: Hsiao, C., Lahiri, K., Lee, L., Pesaran, M.H. (Eds.), Analysis of Panels and Limited Dependent Variables Models: In Honour of G.S. Maddala. Cambridge University Press, Cambridge.

Ahn, S.C., Lee, Y.H., Schmidt, P. (2001). "GMM estimation of linear panel data models with time-varying individual effects". Journal of Econometrics 102, 219–255.

Altonji, J.G., Matzkin, R.L. (2001). "Panel data estimators for nonseparable models with endogenous regressors". NBER Working Paper No. t0267.

Anderson, T.W., Hsiao, C. (1981). "Estimation of dynamic models with error components". Journal of the American Statistical Association 76, 598–606.

Andrews, D., Lu, B. (1999). "Consistent model and moment selection criteria for GMM estimation with application to dynamic panel data". Yale Cowles Foundation Discussion Paper No. 1233.

Arellano, M. (1990). "Testing for autocorrelation in dynamic random effects models". Review of Economic Studies 57, 127–134.

Baker, M. (1997). "Growth-rate heterogeneity and the covariance structure of life-cycle earnings". Journal of Labor Economics 15, 338–375.

Baltagi, B.H., Kao, C. (2000). "Nonstationary panels, cointegration in panels and dynamic panels: A survey". Advances in Econometrics 15, 7–51.

Baltagi, B.H., Song, S.H., Jung, B.C. (2002). "A comparative study of alternative estimators for the unbalanced two-way error component regression model". Econometrics Journal 5, 480–493.

Bekker, P.A. (1994). "Alternative approximations to the distribution of instrumental variables estimators". Econometrica 62, 657–682.

Bergstrom, A.R. (1984). "Continuous time stochastic models and issues of aggregation over time". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. II. North-Holland, Amsterdam. Chapter 20.

Blundell, R., Bond, S. (1998). "Initial conditions and moment restrictions in dynamic panel data models". Journal of Econometrics 87, 115–143.

Blundell, R., Bond, S., Windmeijer, F. (2000). "Estimation in dynamic panel data models: Improving on the performance of the standard GMM estimator". In: Baltagi, B. (Ed.), Advances in Econometrics, vol. 15. Elsevier Science, Amsterdam.

Borjas, G.J. (2002). "The wage structure and the sorting of workers into the public sector". NBER Working Paper No. w9313.

Bound, J., Johnson, G. (1992). "Changes in the structure of wages in the 1980s: An evaluation of alternative explanations". American Economic Review 82, 371–392.

Box, G.E.P., Jenkins, G.N. (1970). Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco.

Breitung, J., Lechner, M. (1999). "Alternative GMM methods for nonlinear panel data models". In: Matyas, L. (Ed.), Generalized Method of Moments Estimation. Cambridge University Press, Cambridge.

Cermeno, R. (1999). "Median unbiased estimation in fixed effects dynamic panels". Annales d'Economie et de Statistique 55–56, 351–368.

Chamberlain, G. (1992). "Comment: Sequential moment restrictions in panel data". Journal of Business and Economic Statistics 10, 20–26.

Chamberlain, G., Hirano, K. (1999). "Predictive distributions based on longitudinal earnings data". Annales d'Economie et de Statistique 55–56, 211–242.

Chib, S., Hamilton, B.H. (2002). "Semiparametric Bayes analysis of longitudinal data treatment models". Journal of Econometrics 110, 67–89.

Choi, I. (2002). "Instrumental variables estimation of a nearly nonstationary, heterogeneous error component model". Journal of Econometrics 109, 1–32.

Chumacero, R.A. (2001). "Estimating ARMA models efficiently". Studies in Nonlinear Dynamics and Econometrics 5, 103–114.

Coakley, J., Fuerts, A., Smith, R.P. (2002). "A principal components approach to cross-section dependence in panels". Unpublished manuscript. Birckbeck College, University of London.

Conley, T.G. (1999). "GMM estimation with cross-sectional dependence". Journal of Econometrics 92, 1–45.

Cragg, J.C. (1985). "More efficient estimation in the presence of heteroscedasticity of unknown form". Econometrica 51, 751–763.

Davis, P. (2002). "Estimating multi-way error components models with unbalanced data structures". Journal of Econometrics 106, 67–95.

Dhrymes, P.J. (1986). "Limited dependent variables". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. III. North-Holland, Amsterdam. Chapter 27.

Druska, V., Horrace, W.C. (2003). "Generalized moments estimation for panel data". NBER Working Paper No. t0291.

DuMoucel, W.H., Duncan, G.J. (1983). "Using sample survey weights in multiple regression analysis of stratified samples". Journal of the American Statistical Association 78, 681–700.

Dustmann, C., Rochina-Barrachina, M.E. (2000). "Selection correction in panel data models: An application to labour supply and wages". Discussion Paper No. 162. Institute for the Study of Labor (IZA).

Edwards, T.H., Whalley, J. (2002). "Short and long run decompositions of OECD wage inequality changes". NBER Working Paper No. w9265.

Efron, B., Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman and Hall, New York.

Engle, R.F. (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of the United Kingdom inflation". Econometrica 50, 987–1007.

Fitzenberger, B., MaCurdy, T. (1997). "Estimation of grouped-data models by block bootstrap procedures". Stanford Working Paper.

Flavin, M. (1981). "The adjustment of consumption to changing expectations about future income". Journal of Political Economy 89, 974–1009.

Gottschalk, P., Moffitt, R. (1994). "The growth of earnings instability in the US labor market". Brookings Papers on Economic Activity 2, 217–272.

Hadri, K. (2001). "Testing for stationarity in heterogeneous panel data". Econometrics Journal 1, 1–14.

Hahn, J., Kuersteiner, G. (2002). "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large". Econometrica 70, 1639–1657.

Hall, A.R. (2001). "Generalized method of moments". In: Baltagi, B. (Ed.), Companions to Contemporary Economics. Blackwell, Oxford.

Hall, P. (1994). "Methodology and theory for the bootstrap". In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, Amsterdam. Chapter 39.

Hall, R.E., Mishkin, F.S. (1982). "The sensitivity of consumption to transitory income: Estimates from panel data on households". Econometrica 50, 461–481.

Heckman, J.J., MaCurdy, T.E. (1986). "Labor econometrics". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. III. North-Holland, Amsterdam. Chapter 32.

Heckman, J.J., Lyons, T., Todd, P. (2000). "Understanding black-white wage differentials, 1960–1990". American Economic Review 92, 344–349.

Heckman, J.J., Lochner, L.J., Todd, P.E. (2003). "Fifty years of mincer earnings regressions". NBER Working Paper No. w9732.

Hirano, K. (2002). "Semiparametric Bayesian inference in autoregressive panel data models". Econometrica 70, 781–799.

Hsiao, C., Pesaran, M.H. (1997). "Bayes estimation of short-run coefficients in dynamic panel data models". University of Cambridge, Department of Applied Economics Working Papers, Amalgamated Series No. 9804.

Hsiao, C., Pesaran, M.H., Tahmiscioglu, A.K. (1999). "Bayes estimation of short-run coefficients in dynamic panel data models". In: Hsiao, C., Lahiri, K., Lee, L.F., Pesaran, M.H. (Eds.), Analysis of Panels and Limited Dependent Variables: A Volume in Honour of G.S. Maddala. Cambridge University Press, Cambridge.

Juhn, C., Murphy, K.M., Pierce, B. (1993). "Wage inequality and the rise in returns to skill". Journal of Political Economy 101, 410–442.

Katz, L.F., Autor, D.H. (1999). "Changes in the wage structure and earnings inequality". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3A. North-Holland, Amsterdam. Chapter 26.

Katz, L.F., Murphy, K.M. (1992). "Changes in the wage structure, 1963–1987: Supply and demand factors". Quarterly Journal of Economics 107, 35–78.

Kelejian, H., Prucha, I. (1999). "A generalized moments estimator for the autoregressive parameter in a spatial model". International Economic Review 40, 509–533.

Kniesner, T.J., Li, Q. (2002). "Nonlinearity in dynamic adjustment: Semiparametric estimation of panel labor supply". Empirical Economics 27, 131–148.

Kuersteiner, G.M. (2001). "Optimal instrumental variables estimation for ARMA models". Journal of Econometrics 104, 395–405.

Kumar, S., Ullah, A. (2000). "Semiparametric varying parametric panel data models: An application to estimation of speed of convergence". In: Fomby, T.B., Hill, R.C. (Eds.), Advances in Econometrics, vol. 14. JAI Press, Stamford, CT.

Kyriazidou, E. (1997). "Estimation of a panel data sample selection model". Econometrica 65, 1335–1364.

Lancaster, T., Aiyar, S. (2000). "Econometric analysis of dynamic panel data models: A growth theory example". In: Bunzel, H. (Ed.), Panel Data and Structural Labour Market Models. Elsevier Science, Amsterdam.

MaCurdy, T. (1982b). "The use of time series processes to model the error structure of earnings in a longitudinal data analysis". Journal of Econometrics 18, 83–114.

MaCurdy, T. (1985). "A guide to applying time series models to panel data". (First draft, NBER Working Paper No. w0624, 1981), NORC Discussion Paper 86-12.

MaCurdy, T., Morz, T. (1989). "Measuring macroeconomic shifts in wages from cohort specifications". Stanford Working Paper (revised 1995).

Maddala, G.S. (1994). "To pool or not to pool: That is the question". In: Maddala, G.S. (Ed.), Econometric Methods and Applications, vol. 1. Ashgate Press, Broomfield, VT.

Mark, N.C., Sul, D. (1999). "A computationally simple cointegration vector estimator for panel data". Mimeo. Ohio State University.

Mark, N.C., Ogaki, M., Sul, D. (2003). "Dynamic seemingly unrelated cointegrating regression". NBER Working Paper No. t0292.

Matzkin, R.L. (1999). "Nonparametric estimation of nonadditive random functions". Mimeo. Northwestern University.

Mincer, J. (1958). "Investment in human capital and personal income distribution". Journal of Political Economy 66, 281–302.

Moffitt, R., Gottschalk, P. (1994). "Trends in the autocovariance structure of earnings in the US: 1969–1987". Mimeo. Department of Economics, Brown University.

Moon, H.R., Phillips, P. (1999). "Maximum likelihood estimation in panels with incidental trends". Yale Cowles Foundation Discussion Paper No. 1246.
Moon, R.H., Perron, B. (2000). "The seemingly unrelated dynamic cointegration regression model and testing for purchasing power parity". Mimeo. University of Southern California.
Moretti, E. (2002). "Estimating the social return to higher education: Evidence from longitudinal and repeated cross-sectional data". NBER Working Paper No. w9108.
Movshuk, O. (2003). "Does the choice of detrending method matter in demand analysis?". Japan and the World Economy 15, 341–359.
Munkin, M.K., Trivendi, P.K. (2003). "Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: An application to the demand for healthcare". Journal of Econometrics 114, 197–200.
Murphy, K.M., Welch, F. (1992a). "Empirical age-earnings profiles". Journal of Labor Economics 8, 202–229.
Murphy, K.M., Welch, F. (1992b). "The structure of wages". Quarterly Journal of Economics 107, 215–326.
Nandram, B., Petruccelli, J.D. (1997). "A Bayesian analysis of autoregressive time series panel data". Journal of Business and Economic Statistics 15, 328–334.
Neal, D. (2002). "The measured black–white wage gap among women is too small". NBER Working Paper No. w9133.
Nerlove, M. (1967). "Distributed lags and unobserved components in economic time series". In: Ten Economic Studies in the Tradition of Irving Fisher. John Wiley and Sons, New York. Chapter 6.
Nevo, A. (2001). "Using weights to adjust for sample selection when auxiliary information is available". NBER Working Paper No. t0275.
Nevo, A. (2002). "Sample selection and information-theoretic alternatives to GMM". Journal of Econometrics 107, 149–157.
Newbold, P. (1978). "Feedback induced by measurement errors". International Economic Review 19, 787–791.
Park, J.Y., Ogaki, M. (1991). "Seemingly unrelated canonical cointegrating regressions". Rochester Center for Economic Research Working Paper No. 280.
Pesaran, H.M. (2003). "Estimation and inference in large heterogenous panels with cross section dependence". University of Cambridge CESifo Working Paper Series No. 869.
Pfeffermann, D. (1993). "The role of sampling weights when modeling survey data". International Statistical Review 61, 317–337.
Phillips, P.C.B., Moon, H.R. (1999). "Linear regression limit theory for nonstationary panel data". Econometrica 67, 1057–1112.
Phillips, P.C.B., Sul, D. (2000). "Dynamic panel estimation and homogeneity testing under cross section dependence". Cowles Foundation Discussion Paper No. 1362.
Phillips, P.C.B., Sul, D. (2002). "Dynamic panel estimation and homogeneity testing under cross section dependence". Mimeo. Yale University.
Phillips, R.F. (2003). "Estimation of a stratified error-components model". International Economic Review 44, 501–521.
Ridder, G. (1992). "An empirical evaluation of some models for nonrandom attrition in panel data". Structural Change and Economic Dynamics 3, 337–355.
Robertson, D., Symons, J. (2000). "Factor residuals in SUR regressions: Estimating panels allowing for cross sectional correlation". Unpublished manuscript. Faculty of Economics and Politics, University of Cambridge.
Saikkonen, P. (1991). "Asymptotically efficient estimation of cointegration regressions". Econometric Theory 7, 1–21.
Sentana, E. (1995). "Quadratic ARCH models". Review of Economic Studies 62, 639–661.
Sims, C. (1974). "Distributed lags". In: Intriligator, M., Kendrick, D. (Eds.), Frontiers of Quantitative Economics, vol. II, pp. 289–332.
Solon, G. (1999). "Intergenerational mobility in the labor market". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3A. North-Holland, Amsterdam. Chapter 29.

Ullah, A., Roy, N. (1998). "Nonparametric and semiparametric econometrics of panel data". In: Ullah, A., Giles, D.E.A. (Eds.), Handbook of Applied Economic Statistics. Dekker, New York. Chapter 17.

Wallis, K. (1977). "Multiple time series analysis and the final form of econometric models". Econometrica 45, 1481–1498.

Wooldridge, J.M. (1999). "Asymptotic properties of weighted M-estimators for variable probability samples". Econometrica 67, 1385–1406.

Wooldridge, J.M. (2001). "Asymptotic properties of weighted M-Estimators for standard stratified samples". Econometric Theory 17, 451–470.

Wooldridge, J.M. (2002). Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge.

Zellner, A., Palm, F. (1974). "Time series analysis and simultaneous equation models". Journal of Econometrics 2, 17–54.

Zellner, A., Palm, F. (1975). "Time series and structural analysis of monetary models of the US economy". Sankya, Series C 37, 12–56. Part 2.

Ziliak, J.P. (1997). "Efficient estimation with panel data when instruments are predetermined: An empirical comparison of moment-condition estimators". Journal of Business Economics and Statistics 15, 419–431.

Chapter 63

ECONOMETRIC TOOLS FOR ANALYZING MARKET OUTCOMES*

DANIEL ACKERBERG
UCLA, USA

C. LANIER BENKARD
Stanford University, USA

STEVEN BERRY
Yale University, USA

ARIEL PAKES
Harvard University, USA

Contents

* The authors thank their NSF grants for financial support.

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06063-1

Abstract

This paper outlines recently developed techniques for estimating the primitives needed to empirically analyze equilibrium interactions and their implications in oligopolistic markets. It is divided into an introduction and three sections; a section on estimating demand functions, a section on estimating production functions, and a section on estimating "dynamic" parameters (parameters estimated through their implications on the choice of controls which determine the distribution of future profits).

The introduction provides an overview of how these primitives are used in typical I.O. applications, and explains how the individual sections are structured. The topics of the three sections have all been addressed in prior literature. Consequently each section begins with a review of the problems I.O. researchers encountered in using the prior approaches. The sections then continue with a fairly detailed explanation of the recent techniques and their relationship to the problems with the prior approaches. Hopefully the detail is rich enough to enable the reader to actually program up a version of the techniques and use them to analyze data. We conclude each section with a brief discussion of some of the problems with the more recent techniques. Here the emphasis is on when those problems are likely to be particularly important, and on recent research designed to overcome them when they are.

Keywords

demand estimation, production function estimation, dynamic estimation, strategic interactions, equilibrium outcomes

JEL classification: C1, C3, C5, C7, L1, L4, L5

Recent complementary developments in computing power, data availability, and econometric technique have led to rather dramatic changes in the way we do empirical analysis of market interactions. This paper reviews a subset of the econometric techniques that have been developed. The first section considers developments in the estimation of demand systems, the second considers developments in the estimation of production functions, and the third is on dynamic estimation, in particular on estimating the costs of investment decisions (where investments are broadly interpreted as any decision which affects future, as well as perhaps current, profits).

These are three of the primitives that are typically needed to analyze market interactions in imperfectly competitive industries. To actually do the analysis, that is to actually unravel the causes of historical events or predict the impact of possible policy changes, we need more information than is contained in these three primitives. We would also need to know the appropriate notion of equilibrium for the market being analyzed, and provide a method of selecting among equilibria if more than one of them were consistent with our primitives and the equilibrium assumptions. Though we will sometimes use familiar notions of equilibrium to develop our estimators, this paper does not explicitly consider either the testing of alternative equilibrium assumptions or the issue of how one selects among multiple equilibria. These are challenging tasks which the profession is just now turning to.

For each of the three primitives we do analyze, we begin with a brief review of the dominant analytic frameworks circa 1990 and an explanation of why those frameworks did not suffice for the needs of modern Industrial Organization. We then move on to recent developments. Our goal here is to explain how to use the recently developed techniques and to help the reader identify problems that might arise when they are used. Each of the three sections have a different concluding subsection.

There have been a number of recent papers which push the demand estimation literature in different directions, so we conclude that section with a brief review of those articles and why one might be interested in them. The section on production function concludes with a discussion of the problems with the approach we outline, and some suggestions for overcoming them (much of this material is new). The section on the costs of investments, which is our section on "dynamics", is largely a summary and integration of articles that are still in various stages of peer review; so we conclude here with some caveats to the new approaches.

We end this introduction with an indication of the ways Industrial Organization makes use of the developments outlined in each of the three sections of the paper. This should direct the researcher who is motivated by particular substantive issues to the appropriate section of the paper. Each section is self-contained, so the reader ought to be able to read any one of them in isolation.

Demand systems are used in several contexts. First demand systems are *the* major tool for comparative static analysis of any change in a market that does not have an immediate impact on costs (examples include the likely effects of mergers, tax changes, etc.). The static analysis of the change usually assumes a mode of competition (almost always either Nash in prices or in quantities) and either has cost data, or more frequently

estimates costs from the first order conditions for a Nash equilibrium. For example, in a Nash pricing (or Bertrand) equilibrium with single product firm, price equals marginal cost plus a markup. The markup can be computed as a function of the estimated demand parameters, so marginal costs can be estimated as price minus this markup. Given marginal costs, demand, and the Nash pricing assumption the analyst can compute an equilibrium under post change conditions (after the tax or the merger). Assuming the computed equilibrium is the equilibrium that would be selected, this generates the predictions for market outcomes after the change. If the analyst uses the pre-change data on prices to estimate costs, the only primitive required for this analysis is the demand function and the ownership pattern of the competing products (which is usually observed).

A second use of demand systems is to analyze the effect of either price changes or new goods on consumer welfare. This is particularly important for the analysis of markets that are either wholly or partially regulated (water, telecommunications, electricity, postage, medicare and medicaid, ...). In this context we should keep in mind that many regulatory decisions are either motivated by nonmarket factors (such as equity considerations), or are politically sensitive (i.e. usually either the regulators or those who appointed them are elected). As a result the analyst often is requested to provide a distribution of predicted demand and welfare impacts across demographic, income and location groups. For this reason a "representative agent" demand system simply will not do.

The use of demand systems to analyze welfare changes is also important in several other contexts. The "exact" consumer price index is a transform of the demand system. Thus ideally we would be using demand systems to construct price indices also (and there is some attempt by the BLS research staff to construct experimental indexes in this way). Similarly the social returns to (either publicly or privately funded) research or infrastructure investments are often also measured with the help of demand systems.

Yet a third way in which demand systems are important to the analysis of I.O. problems is that some of them can be used to approximate the likely returns to potential new products. Demand systems are therefore an integral part of the analysis of product placement decisions, and more generally, for the analysis of the dynamic responses to any policy or environmental change. Finally the way in which tastes are formed, and the impacts of advertising on that process, are problems of fundamental interest to I.O. Unfortunately these are topics we will not address in the demand section of this paper. Our only consolation is the hope that the techniques summarized here will open windows that lead to a deeper understanding of these phenomena.

Production or cost functions are a second primitive needed for comparative static analysis. However partly because product specific cost data are not available for many markets, the direct estimation of cost functions has not been an active area of research lately. There are exceptions, notably some illuminating studies of learning by doing [see Benkard (2000) and the literature cited there], but not many of them.

What has changed in the past decade and a half is that researchers have gained access to a large number of plant (sometimes firm) level data sets on production inputs and outputs (usually the market value of outputs rather than some measure of the physical quantity of the output). This data, often from various census offices, has stimulated renewed interest in production function estimation and the analysis of productivity. The data sets are typically (though not always) panels, and the availability of the data has focused attention on a particular set of substantive and technical issues.

Substantively, there has been a renewal of interest in measuring productivity and gauging how some of the major changes in the economic environment that we have witnessed over the past few decades affect it. This includes studies of the productivity impacts of deregulation, changes in tariff barriers, privatization, and broad changes in the institutional environment (e.g. changes in the legal system, in health care delivery, etc.). The micro data has enabled this literature to distinguish between the impacts of these changes on two sources of growth in aggregate productivity: (i) growth in the productivity of individual establishments, and (ii) growth in industry productivity resulting from reallocating more of the output to the more productive establishments (both among continuing incumbents, and between exitors and new entrants). Interestingly, the prior literature on productivity was also divided in this way. One part focused on the impacts of investments, in particular of research and infrastructure investments, on the productive efficiency of plants. The other focused on the allocative efficiency of different market structures and the impacts of alternative policies on that allocation (in particular of merger and monopoly policy).

From an estimation point of view, the availability of large firm or plant level panels and the desire to use them to analyze the impacts of major changes in the environment has renewed interest in the analysis of the effects of simultaneity (endogeneity of inputs) and selection (endogeneity of attrition) on parameter estimates. The data made clear that there are both: (i) large differences in measured "productivity" across plants (no matter how one measures productivity) and that these differences are serially correlated (and hence likely to effect input choices), and (ii) large sample attrition and addition rates in these panels [see Dunne, Roberts and Samuelson (1988) and Davis and Haltiwanger (1992) for some of the original work on US manufacturing data]. Moreover, the changes in the economic environment that we typically analyze had different impacts on different firms. Not surprisingly, the firms that were positively impacted by the changes tended to have disproportionate growth in their inputs, while those that it affected negatively tended to exhibit falling input demand, and not infrequently, to exit.

The traditional corrections for both simultaneity and selection, corrections based largely on simple statistical models (e.g. use of fixed effect and related estimators for simultaneity, and the use of the propensity score for selection) were simply not rich enough to account for the impacts of such major environmental changes. So the literature turned to simultaneity and selection corrections based on economic models of input and exit choices. The section of this chapter on production functions deals largely with these latter models. We first review the new procedures emphasizing the assumptions

they use, and then provide suggestions for amending the estimators for cases where those assumptions are suspect.

The last section of the paper deals explicitly with dynamic models. Despite a blossoming empirical literature on the empirical analysis of static equilibrium models, there has been very little empirical work based on dynamic equilibrium models to date. The I.O. literature's focus on static settings came about not because dynamics were thought to be unimportant to the outcomes of interest. Indeed it is easy to take any one of the changes typically analyzed in static models and make the argument that the dynamic implications of the change might well overturn their static effects. Moreover, there was a reasonable amount of agreement among applied researchers that the notion of Markov perfect equilibrium provided a rich enough framework for the analysis of dynamics in oligopolistic settings.

The problem was that even given this framework the empirical analysis of the dynamic consequences of the changes being examined was seen as too difficult a task to undertake. In particular, while some of the parameters needed to use the Markov perfect framework to analyze dynamic games could be estimated without imposing the dynamic equilibrium conditions, some could not. Moreover, until very recently the only available methods for estimating these remaining parameters were extremely burdensome, in terms of both computation and researcher time.

The computational complexity resulted from the need to compute the continuation values to the dynamic game in order to estimate the model. The direct way of obtaining continuation values was to compute them as the fixed point to a functional equation, a high order computational problem. Parameter values were inferred from observed behavior by computing the fixed point that determines continuation values at different trial parameter values, and then searching for the parameter value that makes the behavior implied by the continuation values "as close as possible" to the observed data. This "nested fixed point" algorithm is extremely computationally burdensome; the continuation values need to be computed many times and each time they are computed we need to solve the fixed point.

A recent literature in industrial organization has developed techniques that substantially reduce the computational and programming burdens of using the implications of dynamic games to estimate the parameters needed for subsequent applied analysis. That literature requires some strong assumptions, but delivers estimating equations which have simple intuitive explanations and are easy to implement.

Essentially the alternative techniques deliver different semiparametric estimates of continuation values. Conditional on a value of the parameter vector, these estimated continuation values are treated as the true continuation values and used to determine optimal policies (these can be entry and exit policies, investments of various forms, or bidding strategies in dynamic auctions). The parameters are estimated by matching the policies that are predicted in this way to the policies that are observed in the data. Note that this process makes heavy use of nonparametric techniques; nonparametric estimates of either policies or values must be estimated at every state observed in the data. Not surprisingly then Monte Carlo evidence indicates that the small sample properties of the

estimators can be quite important in data sets of the size we currently use. This, in turn, both generates preferences for some semiparametric estimators over others, and makes obvious a need for small sample bias correction procedures which, for the most part, have yet to be developed. We now move on to the body of the paper.

1. Demand systems

Demand systems are probably the most basic tool of empirical Industrial Organization. They summarize the demand preferences that determine the incentives facing producers. As a result some form of demand system has to be estimated before one can proceed with a detailed empirical analysis of pricing (and/or production) decisions, and, consequently of the profits and consumer welfare likely to be generated by the introduction of new goods.

Not long ago graduate lectures on demand systems were largely based on "representative agent" models in "product" space (i.e. the agent's utility was defined on the product per se rather than on the characteristics of the product). There were a number of problems with this form of analysis that made if difficult to apply in the context of I.O. problems. We begin with an overview of those problems, and the "solutions" that have been proposed to deal with them.

Heterogeneous agents and simulation

First, almost all estimated demand systems were based on market level data: they would regress quantity purchased on (average) income and prices. There were theoretical papers which investigated the properties of market level demand systems obtained by explicitly aggregating up from micro models of consumer choices [including a seminal paper by Houthakker (1955)]. However we could not use their results to structure estimation on market level data without imposing unrealistic *a priori* assumptions on the distribution of income and "preferences" (or its determinants like size, age, location, etc.) across consuming units.

Simulation estimators, which Pakes (1986) introduced for precisely this problem, i.e. to enable one to use a micro behavioral model with heterogeneity among agents to structure the empirical analysis of aggregate data, have changed what is feasible in this respect. We can now aggregate up from the *observed* distribution of consumer characteristics and any functional form that we might think relevant. That is we allow different consumers to have different income, age, family size, and/or location of residence. We then formulate a demand system which is conditional on the consumer's characteristics and a vector of parameters which determines the relationship between those characteristics and preferences over products (or over product characteristics). To estimate those parameters from market level data we simply

- draw vectors of consumer characteristics from the distribution of those characteristics in the market of interest (in the US, say from the March CPS),

- determine the choice that each of the households drawn would make for a given value of the parameter vector,
- aggregate those choices into a prediction for aggregate demand conditional on the parameter vector, and
- employ a search routine that finds the value of that parameter vector which makes these aggregate quantities as close as possible to the observed market level demands.

The ability to obtain aggregate demand from a distribution of household preferences has had at least two important impacts on demand analysis. First it has allowed us to use the same framework to study demand in different markets, or in the same market at different points in time. A representative agent framework might generate a reasonable approximation to a demand surface in a particular market. However there are often large differences in the distribution of income and other demographic characteristics across markets, and these in turn make an approximation which fits well in one market do poorly in others.

For example, we all believe (and virtually all empirical work indicates) that the impact of price depends on income. Our micro model will therefore imply that the price elasticity of a given good depends on the density of the income distribution among the income/demographic groups attracted to that good. So if the income distribution differed across regional markets, and we used an aggregate framework to analyze demand, we would require different price coefficients for each market. Table 1 provides some data on the distribution of the income distribution across US counties (there are about three thousand counties in the US). It is clear that the income distribution differs markedly across these "markets"; the variance being especially large in the high income groups (the groups which purchase a disproportionate share of goods sold).

Table 1
Cross county differences in household income

Income group (thousands)	Fraction of US population in income group	Distribution of fraction over counties	
		Mean	Std. dev.
0–20	0.226	0.289	0.104
20–35	0.194	0.225	0.035
35–50	0.164	0.174	0.028
50–75	0.193	0.175	0.045
75–100	0.101	0.072	0.033
100–125	0.052	0.030	0.020
125–150	0.025	0.013	0.011
150–200	0.022	0.010	0.010
200+	0.024	0.012	0.010

Source: From Pakes (2004).

A heterogenous agent demand model with an interaction between price and income uses the available information on differences in the distribution of income to combine the information from different markets. This both enables us to obtain more precise parameter estimates, and provides a tool for making predictions of likely outcomes in new markets.

The second aspect of the heterogenous agent based systems that is intensively used is its ability to analyze the distributional impacts of policies or environmental changes that affect prices and/or the goods marketed. These distributional effects are often of primary concern to both policy makers and to the study of related fields (e.g. the study of voting patterns in political economy, or the study of tax incidence in public finance).

The too many parameters and new goods problems

There were at least two other problems that appeared repeatedly when we used the earlier models of demand to analyze Industrial Organization problems. They are both a direct result of positing preferences directly on products, rather than on the characteristics of products.

1. Many of the markets we wanted to analyze contained a large number of goods that are substitutes for one another. As a result when we tried to estimate demand systems in product space we quickly ran into the "too many parameters problem". Even a (log) linear demand system in product space for J products requires estimates of on the order of J^2 parameters (J price and one income coefficient in the demand for every one of the J products). This was often just too many parameters to estimate with the available data.
2. Demand systems in product space do not enable the researcher to analyze demand for new goods prior to their introduction.

Gorman's polar forms [Gorman (1959)] for multi-level budgeting were an ingenious attempt to mitigate the too many parameter problem. However they required assumptions which were often unrealistic for the problem at hand. Indeed typically the grouping procedures used empirically paid little attention to accommodating Gorman's conditions. Rather they were determined by the policy issue of interest. As a result one would see demand systems for the same good estimated in very different ways with results that bore no relationship to each other.[1] Moreover, the reduction in parameters obtained from multilevel budgeting was not sharp enough to enable the kind of flexibility needed for many I.O. applications [though it was for some, see for e.g. Hausman (1996) and the literature cited there].

[1] For example, it was not uncommon to see automobile demand systems that grouped goods into imports and domestically produced in studies where the issue of interest involved tariffs of some form, and alternatively by gas mileage in studies where the issue of interest was environmental or otherwise related to fuel consumption. Also Gorman's results were of the "if and only if" variety; one of his two sets of conditions were necessary if one is to use multi-level budgeting. For more detail on multi-level budgeting see Deaton and Muellbauer (1980).

The new goods problem was central to the dynamics of analyzing market outcomes. That is in order to get any sort of idea of the incentives for entry in differentiated product markets, we need to be able to know something about the demand for a good which had not yet been introduced. This is simply beyond the realm of what product based demand systems can do. On the other hand entry is one of the basic dynamic adjustment mechanisms in Industrial Organization, and it is hard to think of say, the likely price effects of a merger,[2] or the longer run effects of an increase in gas prices, without some way of evaluating the impacts of those events on the likelihood of entry.

The rest of this section of the paper will be based on models of demand that posit preferences on the characteristics of products rather than on products themselves. We do not, however, want to leave the reader with the impression that demand systems in product based, in particular product space models that allow for consumer heterogeneity, should not be used. If one is analyzing a market with a small number of products, and if the issue of interest does not require an analysis of the potential for entry, then it may well be preferable to use a product space system. Indeed all we do when we move to characteristic space is to place restrictions on the demand systems which could, at least in principle, be obtained from product space models. On the other hand these restrictions provide a way of circumventing the "too many parameter" and "new goods" problems which has turned out to be quite useful.

1.1. Characteristic space: The issues

In characteristics space models

- Products are bundles of characteristics.
- Preferences are defined on those characteristics.
- Each consumer chooses a bundle that maximizes its utility. Consumers have different relative preferences (usually just marginal preferences) for different characteristics, and hence make different choices.
- Simulation is used to obtain aggregate demand.

Note first that in these models the number of parameters required to determine aggregate demand is *independent* of the number of products per se; all we require is the joint distribution of preferences over the characteristics. For example, if there were five important characteristics, and preferences over them distributed joint normally, twenty parameters would determine the own and cross price elasticities for all products (no matter the number of those products). Second, once we estimate those parameters, if we specify a new good as a different bundle of characteristics then the bundles currently in existence, we can predict the outcomes that would result from the entry of the new good

[2] Not surprisingly, then, directly after explaining how they will analyze the price effects of mergers among incumbent firms, the US merger guidelines [DoJ (1992)] remind the reader that the outcome of the analysis might be modified by an analysis of the likelihood of entry. Though they distinguish between different types of potential entrants, their guidelines for evaluating the possibility of entry remain distinctly more *ad hoc* then the procedures for analyzing the initial price changes.

by simply giving each consumer an expanded choice set, one that includes the old and the new good, and recomputing demand in exactly the same way as it was originally computed.[3]

Having stated that, at least in principle, the characteristic space based systems solve both the too many parameter and the new goods problems, we should now provide some caveats. First what the system does is restrict preferences: it only allows two products to be similar to one another through similarities in their characteristics. Below we will introduce unmeasured characteristics into the analysis, but the extent to which unmeasured characteristics have been used to pick up similarities in tastes for different products is very limited. As a result if the researcher does not have measures of the characteristics that consumers care about when making their purchase decisions, the characteristic based models are unlikely to provide a very useful guide to which products are good substitutes for one another. Moreover, it is these substitution patterns that determine pricing incentives in most I.O. models (and as a result profit margins and the incentives to produce new goods).

As for new goods, there is a very real sense in which characteristic based systems can only provide adequate predictions for goods that are not too "new". That is, if we formed the set of all tuples of characteristics which were convex combinations of the characteristics of existing products, and considered a new product whose characteristics are outside of this set, then we would not expect the estimated system to be able to provide much information regarding preferences for the new good, as we would be "trying to predict behavior outside of the sample". Moreover, many of the most successful product introductions are successful precisely because they consist of a tuple of characteristics that is very different than any of the characteristic bundles that had been available before it was marketed (think, for example, of the lap top computer, or the Mazda Miata[4]).

Some background

The theoretical and econometric groundwork for characteristic based demand systems dates back at least to the seminal work of Lancaster (1971) and McFadden (1974, 1981).[5] Applications of the Lancaster/McFadden framework however, increased significantly after Berry, Levinsohn and Pakes (1995) showed how to circumvent two

[3] This assumes that there are no product specific unobservables. As noted below, it is typically important to allow for such unobservables when analyzing demand for consumer products, and once one allows for them we need to account for them in our predictions of demand for new goods. For example, see Berry, Levinsohn and Pakes (2004).

[4] For more detail on just how our predictions would fail in this case see Pakes (1995).

[5] Actually characteristics based models have a much longer history in I.O. dating back at least to Hotelling's (1929) classic article, but the I.O. work on characteristic based models focused more on their implications for product placement rather than on estimating demand systems per se. Characteristic based models also had a history in the price index literature as a loose rational for the use of hedonic price indices; see Court (1939), Griliches (1961), and the discussion of the relationship between hedonics and I.O. equilibrium models in Pakes (2004).

problems that had made it difficult to apply the early generation of characteristic based models in I.O. contexts.

The problems were that

1. The early generation of models used functional forms which restricted cross and own price elasticities in ways which brought into question the usefulness of the whole exercise.
2. The early generation of models did not allow for unobserved product characteristics.

The second problem was first formulated in a clear way by Berry (1994), and is particularly important when studying demand for consumer goods. Typically these goods are differentiated in many ways. As a result even if we measured all the relevant characteristics we could not expect to obtain precise estimates of their impacts. One solution is to put in the "important" differentiating characteristics *and* an unobservable, say ξ, which picks up the aggregate effect of the multitude of characteristics that are being omitted. Of course, to the extent that producers know ξ when they set prices (and recall ξ represents the effect of characteristics that are known to consumers), goods that have high values for ξ will be priced higher in any reasonable notion of equilibrium.

This produces an analogue to the standard simultaneous equation problem in estimating demand systems in the older demand literature; i.e. prices are correlated with the disturbance term. However in the literature on characteristics based demand systems the unobservable is buried deep inside a highly nonlinear set of equations, and hence it was not obvious how to proceed. Berry (1994) shows that there is a unique value for the vector of unobservables that makes the predicted shares exactly equal to the observed shares. Berry, Levinsohn and Pakes (1995) (henceforth BLP) provide a contraction mapping which transforms the demand system into a system of equations that is linear in these unobservables. The contraction mapping is easy to compute, and once we have a system which is linear in the disturbances we can again use instruments, or any of the other techniques used in more traditional endogeneity problems, to overcome this "simultaneity problem".

The first problem, that is the use of functional forms which restricted elasticities in unacceptable ways, manifested itself differently in different models and data sets. The theoretical I.O. literature focussed on the nature competition when there was one dimension of product competition. This could either be a "vertical" or quality dimension as in Shaked and Sutton (1982) or a horizontal dimension, as in Salop (1979) [and in Hotelling's (1929) classic work]. Bresnahan (1981), in his study of the automobile demand and prices, was the first to bring this class of models to data. One (of several) conclusions of the paper was that a one-dimensional source of differentiation among products simply was not rich enough to provide a realistic picture of demand: in particular it implied that a particular good only had a nonzero cross price elasticity with its two immediate neighbors (for products at a corner of the quality space, there was only one neighbor).

McFadden himself was quick to point out the "IIA" (or independence of irrelevant alternatives) problem of the logit model he used. The simplest logit model, and the one

that had been primarily used when only aggregate data was available (data on quantities, prices, and product characteristics), has the utility of the ith consumer for the jth product defined as

$$U_{i,j} = x_j\beta + \epsilon_{i,j},$$

where the x_j are the characteristics of product j (including the unobserved characteristic and price) and the $\{\epsilon_{i,j}\}$ are independent (across both j for a given i and across i for a given j) identically distributed random variables.[6] Thus $x_j\beta$ is the mean utility of product j and $\epsilon_{i,j}$ is the individual specific deviation from that mean.

There is a rather extreme form of the IIA problem in the demand generated by this model. The model implies that the distribution of a consumer's preferences over products *other than* the product it bought, does not depend on the product it bought. One can show that this implies the following:

- Two agents who buy different products are equally likely to switch to a particular third product should the price of their product rise. As a result two goods with the same shares have the same cross price elasticities with any other good (cross price elasticities are a multiple of $s_j s_k$, where s_j is the share of good j). Since both very high quality goods with high prices and very low quality goods with low prices have low shares, this implication is inconsistent with basic intuition.
- Since there is no systematic difference in the price sensitivities of consumers attracted to the different goods, own price derivatives only depends on shares $(\partial s/\partial p) = -s(1-s)$. This implies that two goods with same share must have the same markup in a single product firm "Nash in prices" equilibrium, and once again luxury and low quality goods can easily have the same shares.

No data will ever change these implications of the two models. If your estimates do not satisfy them, there is a programming error, and if your estimates do satisfy them, we are unlikely to believe the results.

A way of ameliorating this problem is to allow the coefficients on x to be individual-specific. Then, when we increase the price of one good the consumers who leave that good have very particular preferences, they were consumers who preferred the x's of that good. Consequently they will tend to switch to another good with similar x's generating exactly the kind of substitution patterns that we expect to see. Similarly, now consumers who chose high priced cars will tend to be consumers who care less about price. Consequently less of them will substitute from the good they purchase for any given price increase, a fact which will generate lower price elasticities and a tendency for higher markups on those goods.

[6] In the pure logit, they have a double exponential distribution. Though this assumption was initially quite important, it is neither essential for the argument that follows, nor of as much importance for current applied work. Its original importance was due to the fact that it implied that the integral that determined aggregate demand had a closed form, a feature which receded in importance as computers and simulation techniques improved.

This intuition also makes it clear how the IIA problem was ameliorated in the few studies which had micro data (data which matched individual characteristics to the products those individuals chose), and used it to estimate a micro choice model which was then explicitly aggregated into an aggregate demand system. The micro choice model interacted *observed* individual and product characteristics, essentially producing individual specific β's in the logit model above. The IIA problem would then be ameliorated to the extent that the individual characteristic data captured the differences in preferences for different x-characteristics across households. Unfortunately many of the factors that determine different households preferences for different characteristics are typically not observed in our data sets, so without allowing for unobserved as well as observed sources of differences in the β, estimates of demand systems typically retain many reflections of the IIA problem as noted above; see, in particular Berry, Levinsohn and Pakes (2004) (henceforth MicroBLP) and the literature cited there.

The difficulty with allowing for individual specific coefficients on product characteristics in the aggregate studies was that once we allowed for them the integral determining aggregate shares was not analytic. This lead to a computational problem; it was difficult to find the shares predicted by the model conditional on the model's parameter vector. This, in turn, made it difficult, if not impossible, to compute an estimator with desirable properties. Similarly in micro studies the difficulty with allowing for unobserved individual specific characteristics that determined the sensitivity of individuals to different product characteristics was that once we allowed for them the integral determining individual probabilities was not analytic. The literature circumvented these problems as did Pakes (1986), i.e. by substituting simulation for integration, and then worried explicitly about the impact of the simulation error on the properties of the estimators [see Berry, Linton and Pakes (2004) and the discussion below].

1.2. Characteristic space: Details of a simple model

The simplest characteristic based models assumes that each consumer buys at most one unit of one of the differentiated goods. The utility from consuming good j depends on the characteristics of good j, as well as on the tastes (interpreted broadly enough to include income and demographic characteristics) of the household. Heterogenous households have different tastes and so may choose different products.

The utility of consumer (or household) i for good j in market (or time period) t if it purchases the jth good is

$$u_{ijt} = U(\tilde{x}_{jt}, \xi_{jt}, z_{it}, \nu_{it}, y_{it} - p_{jt}, \theta), \tag{1}$$

where $\tilde{x}_{jt}$ is a K-dimensional vector of observed product characteristics other than price, p_{jt} is the price of the product, ξ_{jt} represents product characteristics unobserved to the econometrician, z_{it} and ν_{it} are vectors of observed and unobserved (to the econometrician) sources of differences in consumer tastes, y_{it} is the consumer's income, and θ is a vector of parameters to be estimated. When we discuss decisions within a single market, we will often drop the t subscript.

Note that the "partial equilibrium" nature of the problem is incorporated into the model by letting utility depend on the money available to spend outside of this market $(y_i - p_j)$. In many applications, the expenditure in other markets is not explicitly modelled. Instead, y_i is subsumed into either ν_i or z_i and utility is modelled as depending explicitly on price, so that utility is

$$u_{ij} = U(\tilde{x}_j, \xi_j, z_i, \nu_i, p_j, \theta). \tag{2}$$

The consumer chooses one of j products and also has the $j = 0$ choice of not buying any of the goods (i.e. choosing the "outside option"). Denote the utility of outside good as

$$u_{i0} = U(\tilde{x}_0, \xi_0, z_i, \nu_i, \theta), \tag{3}$$

where $\tilde{x}_0$ could either be a vector of "characteristics" of the outside good, or else could be an indicator for the outside good that shifts the functional form of U (because the outside good may be difficult to place in the same space of product characteristics as the "inside" goods). The existence of the outside option allows us to model aggregate demand for the market's products; in particular it allows market demand to decline if all within-market prices rise.

The consumer makes the choice that gives the highest utility. The probability of that product j is chosen is then the probability that the unobservables ν are such that

$$u_{ij} > u_{ir}, \quad \forall r \neq j. \tag{4}$$

The demand system for the industry's products is obtained by using the distribution of the (z_i, ν_i) to sum up over the values for these variables that satisfy the above condition in the market of interest.

Note that, at least with sufficient information on the distribution of the (z_i, ν_i), the same model can be applied when: only market level data are available, when we have micro data which matches individuals to the choices they make, when we have stratafied samples or information on the total purchases of particular strata, or with any combination of the above types of data. In principal at least, this should make it easy to compare different studies on the same market, or to use information from one study in another.

Henceforth we work with the linear case of the model in Equations (2) and (3). Letting $x_j = (\tilde{x}_j, p_j)$, that model can be written as

$$U_{ij} = \Sigma_k x_{jk}\theta_{ik} + \xi_j + \epsilon_{ij}, \tag{5}$$

with

$$\theta_{ik} = \bar{\theta}_k + \theta_k^{\mathrm{o}\prime} z_i + \theta_k^{\mathrm{u}\prime} \nu_i,$$

where the "o" and "u" superscripts designate the interactions of the product characteristic coefficients with the observed and the unobserved individual attributes, and it is understood that $x_{i0} \equiv 1$.

We have not written down the equation for $U_{i,0}$, i.e. for the outside alternative, because we can add an individual specific constant term to each choice without changing

the order of preferences over goods. This implies we need a normalization and we chose $U_{i,0} = 0$ (that is we subtract $U_{i,0}$ from each choice). Though this is notationally convenient we should keep in mind that the utilities from the various choices are now actually the differences in utility between the choice of the particular good and the outside alternative.[7]

Note also that we assume a single unobservable product characteristic, i.e. $\xi_j \in \mathcal{R}$, and its coefficient does not vary across consumers. That is, if there are multiple unobservable characteristics then we are assuming they can be collapsed into a single index whose form does not vary over consumers. This constraint is likely to be more binding were we to have data that contained multiple choices per person [see, for example Heckman and Snyder (1997)].[8] Keep in mind, however, that any reasonable notion of equilibrium would make p_j depend on ξ_j (as well as on the other product characteristics).

The only part of the specification in (5) we have not explained are the $\{\epsilon_{ij}\}$. They represent unobserved sources of variation that are independent across individuals for a given product, and across products for a given individual. In many situations it is hard to think of such sources of variation, and as a result one might want to do away with the $\{\epsilon_{ij}\}$. We show below that it is possible to do so, and that the model without the $\{\epsilon_{ij}\}$ has a number of desirable properties. On the other hand it is computationally convenient to keep the $\{\epsilon_{ij}\}$, and the model without them is a limiting case of the model with them (see below), so we start with the model in (5). As is customary in the literature, we will assume that the $\{\epsilon_{ij}\}$ are i.i.d. with the double exponential distribution.

Substituting the equation which determines θ_{ik} into the utility function in (5) we have

$$U_{ij} = \delta_j + \Sigma_{kr} x_{jk} z_{ir} \theta^{\mathrm{o}}_{rk} + \Sigma_{kl} x_{jk} \nu_{il} \theta^{\mathrm{u}}_{kl} + \epsilon_{ij}, \tag{6}$$

where

$$\delta_j = \Sigma_k x_{jk} \bar{\theta}_k + \xi_j.$$

Note that the model has two types of interaction terms between product and consumer characteristics: (i) interactions between observed consumer characteristics (the z_i) and product characteristics (i.e. $\Sigma_{kr} x_{jk} z_{ir} \theta^{\mathrm{o}}_{rk}$), and (ii) interactions between unobserved consumer characteristics (the ν_i) and product characteristics (i.e. $\Sigma_{kl} x_{jk} \nu_{il} \theta^{\mathrm{u}}_{kl}$). It is these interactions which generate reasonable own and cross price elasticities (i.e. they are designed to do away with the IIA problem).

[7] We could also multiply each utility by positive constant without changing the order, but we use this normalization up by assuming that the $\epsilon_{i,j}$ are i.i.d. extreme value deviates, see below.

[8] Attempts we have seen to model a random coefficient on the ξ have lead to results which indicate that there was no need for one, see Das, Olley and Pakes (1996).

1.3. Steps in estimation: Product level data

There are many instances in which use of the model in (6) might be problematic, and we come back to a discussion of them below. Before doing so, however, we want to consider how to estimate that model. The appropriate estimation technique depends on the data available and the market being modelled. We begin with the familiar case where only product level demand data is available, and where we can assume that we have available a set of variables w that satisfies $E[\xi|w] = 0$. This enables us to construct instruments to separate out the effect of ξ from that of x in determining shares. The next section considers additional sources of information, and shows how the additional sources of information can be used to help estimate the parameters of the problem. In the section that follows we come back to the "identifying" assumption, $E[\xi|w] = 0$, consider the instruments it suggests, and discuss alternatives.

When we only have product level data all individual characteristics are unobserved, i.e. $z_i \equiv 0$. Typically some of the unobserved individual characteristics, the ν_i will have a known distribution (e.g. income), while some will not. For those that do not we assume that distribution up to a parameter to be estimated, and subsume those parameters into the utility function specification (for example, assume a normal distribution and subsume the mean in $\bar{\theta}_k$ and the standard deviation in θ_k^{u}). The resultant known joint distribution of unobserved characteristics is denoted by $f_\nu(\cdot)$. We now describe the estimation procedure.

The first two steps of this procedure are designed to obtain an estimate of $\xi(\cdot)$ as a function of θ. We then require an identifying assumption that states that at $\theta = \theta_0$, the true value of θ, the distribution of $\xi(\cdot;\theta)$ obeys some restriction. The third step is a standard method of moments step that finds the value of θ that makes the distribution of the estimated $\xi(\cdot,\theta)$ obey that restriction to the extent possible.

STEP I. We first find an approximation to the aggregate shares conditional on a particular value of (δ,θ). As noted by McFadden (1974) the logit assumption implies that, when we condition on the ν_i, we can find the choice probabilities implied by the model in (6) analytically. Consequently the aggregate shares are given by

$$\sigma_j(\theta,\delta) = \int \frac{\exp[\delta_j + \Sigma_{kl} x_{jk}\nu_{il}\theta_{kl}^{\mathrm{u}}]}{1+\sum_q \exp[\delta_q + \Sigma_{kl} x_{qk}\nu_{il}\theta_{kl}^{\mathrm{u}}]} f(\nu)\,\mathrm{d}(\nu). \tag{7}$$

Typically this integral is intractable. Consequently we follow Pakes (1986) and use simulation to obtain an approximation of it. I.e. we take ns pseudo-random draws from $f_\nu(\cdot)$ and compute

$$\sigma_j\left(\theta,\delta,P^{ns}\right) = \sum_{r=1}^{ns} \frac{\exp[\delta_j + \Sigma_{kl} x_{jk}\nu_{ilr}\theta_{kl}^{\mathrm{u}}]}{1+\sum_q \exp[\delta_q + \Sigma_{kl} x_{qk}\nu_{ilr}\theta_{kl}^{\mathrm{u}}]}, \tag{8}$$

where P^{ns} denotes the empirical distribution of the simulation draws. Note that the use of simulation introduces simulation error. The variance of this error decreases with ns

but for given *ns* can be made smaller by using importance sampling or other variance reduction techniques [for a good introduction to these techniques see Rubinstein (1981)]. Below we come back to the question of how the simulation error affects the precision of the parameter estimates.

STEP II. Let the vector of observed shares be $s^n = [s_1^n, \ldots, s_J^n]$, where n denotes the size of the sample from which these shares are calculated (which is often very large). Step II finds the unique values of δ that makes the predicted shares for a given θ and set of simulation draws equal to s^n. BLP show that iterating on the system of equations

$$\delta_j^k(\theta) = \delta_j^{k-1}(\theta) + \ln\left[s_j^n\right] - \ln\left[\sigma_j\left(\theta, \delta^{k-1}, P^{ns}\right)\right] \tag{9}$$

leads to the unique δ that makes $\sigma_j(\theta, \delta, P^{ns}) = s_j^n$ for all j.[9]

Call the fixed point obtained from the iterations $\delta(\theta, s^n, P^{ns})$. The model in (6) then implies that

$$\xi_j\left(\theta, s^n, P^{ns}\right) = \delta\left(\theta, s^n, P^{ns}\right) - \Sigma_k x_{jk}\bar{\theta}_k. \tag{10}$$

I.e. we have solved for the $\{\xi_j\}$ as a function of the parameters, the data, and our simulation draws.

"IDENTIFICATION". An identifying restriction for our model will be a restriction on the distribution of the true ξ, the ξ obtained when we evaluate the above equation at $n = ns = \infty$, that will only be satisfied by $\xi_j(\theta, s^\infty, P^\infty)$ when $\theta = \theta_0$ (but not at other values of θ). Different restrictions may well be appropriate in different applied cases, and we come back to a discussion of possible restrictions below. For now, however, we illustrate by assuming we have a set of instruments, say w that satisfy $E[\xi(\theta_0)|w] = 0$. In that case the third and final step of the algorithm is as follows.

STEP III. Interact $\xi_j(\theta, s^n, P^{ns})$ with function of w and find that value of θ that makes the sample moments as close as possible to zero. I.e. minimize $\|G_{J,n,ns}(\theta)\|$ where

$$G_{J,n,ns}(\theta) = \sum_j \xi_j\left(\theta, s^n, P^{ns}\right) f_j(w). \tag{11}$$

Formal conditions for the consistency and asymptotic normality of this estimator are given in Berry, Linton and Pakes (2004), and provided one accounts for simulation and sampling error in the estimate of the objective function, standard approximations to the limit distribution work [see, for e.g. Pakes and Pollard (1989)]. A few of the properties of this limit distribution are discussed below. For now we want only to note that there is an analytic form for the $\bar{\theta}$ parameters conditional on the θ^{u}; i.e. for the given θ^{u} the solution for $\bar{\theta}$ is given by the standard instrumental variable formula. So the nonlinear search is only over θ^{u}.

[9] Note that one has to recompute the shares at the "new" δ at each iteration. The system of equations is a mapping from possible values of δ into itself. BLP prove that the mapping is a contraction mapping with modulus less that one. The iterations therefore converge geometrically to the unique fixed point of the system.

1.4. Additional sources of information on demand parameters

Often we find that there is not enough information in product level demand data to estimate the entire distribution of preferences with sufficient precision. This should not be surprising given that we are trying to estimate a whole distribution of preferences from just aggregate choice probabilities. Other than functional form, the information that is available for this purpose comes from differences in choice sets across markets or time periods (this allows you to sweep out preferences for given characteristics), and differences in preferences across markets or over time for a fixed choice set (the preferences differences are usually associated with known differences in demographic characteristics). The literature has added information in two ways. One is to add an equilibrium assumption and work out its implications for the estimation of demand parameters, the other is to add data. We now consider each of these in turn.

1.4.1. Adding the pricing equation

There is a long tradition in economics of estimating "hedonic" or reduced form equations for price against product characteristics in differentiated product markets [see, in particular Court (1939) and Griliches (1961)]. Part of the reason those equations were considered so useful, useful enough to be incorporated as correction procedures in the construction of most countries' Consumer Price Indices, was that they typically had quite high R^2's.[10] Indeed, at least in the cross section, the standard pricing equations estimated by I.O. economists have produced quite good fits (i.e. just as the model predicts, goods with similar characteristics tend to sell for similar prices, and goods in parts of the characteristic space with lots of competitors tend to sell for lower prices). Perhaps it is not surprising then that when the pricing system is added to the demand system the precision of the demand parameters estimates tends to improve noticeably (see, for e.g. BLP).

Adding the pricing system from an oligopoly model to the demand system and estimating the parameters of two systems jointly is the analogue of adding the supply equation to the demand equation in a perfectly competitive model and estimating the parameters of those systems jointly. So it should not be surprising that the empirical oligopoly literature itself started by estimating the pricing and demand systems jointly [see Bresnahan (1981)]. On the other hand there is a cost of using the pricing equation. It requires two additional assumptions: (i) an assumption on the nature of equilibrium, and (ii) an assumption on the cost function.

The controversial assumption is the equilibrium assumption. Though there has been some empirical work that tries a subset of the alternative equilibrium assumptions and sees how they fit the data [see, for e.g. Berry, Levinsohn and Pakes (1999) or Nevo

[10] For a recent discussion of the relationship between hedonic regressions and pricing equations with special emphasis on implications for the use of hedonics in the CPI, see Pakes (2004).

(2001)], almost all of it has assumed static profit maximization, no uncertainty, and that one side of the transaction has the power to set prices while the other can only decide whether and what to buy conditional on those prices. There are many situations in which we should expect current prices to depend on likely future profits (e.g.'s include any situation in which demand or cost tomorrow depends on current sales, and/or where there are collusive possibilities; for more discussion see the last section of this chapter). Additionally there are many situations, particularly in markets where vertical relationships are important, where there are a small number of sellers facing a small number of buyers; situations where we do not expect one side to be able to dictate prices to another [for an attempt to handle these situations see Pakes et al. (2006)].

On the other hand many (though not all) of the implications of the results that are of interest will require the pricing assumption anyway, so there might be an argument for using it directly in estimation. Moreover, as we have noted, the cross-sectional distribution of prices is often quite well approximated by our simple assumptions, and, partly as a result, use of those assumptions is often quite helpful in sorting out the relevance of alternative values of θ.

We work with a Nash in prices, or Bertrand, assumption. Assume that marginal cost, to be denoted by mc, is log linear in a set of observables r_{kj} and a disturbance which determines productivity or ω_j, i.e.

$$\ln[\mathrm{mc}_j] = \sum r_{k,j}\theta_k^c + \omega_j. \tag{12}$$

r will typically include product characteristics, input prices and, possibly the quantity produced (if there are nonconstant returns to scale). As a result our demand and cost disturbances (i.e. ξ and ω) will typically be mean independent of some of the components of r but not of others. Also we might expect a positive correlation between ξ and ω since goods with a higher unobserved quality might well cost more to produce.

Since we characteristically deal with multiproduct firms, and our equilibrium assumption is that each firm sets each of its prices to maximize the profits from all of its products conditional on the prices set by its competitors, we need notation for the set of products owned by firm f, say J_f. Then the Nash condition is that firms set each of their prices to maximize $\sum_{j\in J_f}(p_j - C_j(\cdot))Ms_j(\cdot)$, where C_j is total costs. This implies that for $j = 1, \ldots, J$

$$\sigma_j(\cdot) + \sum_{l\in J_f}(p_l - \mathrm{mc}_l)M\frac{\partial\sigma_l(\cdot)}{\partial p_j} = 0. \tag{13}$$

Note that we have added a system of J equations (one for each price) and $R = \dim(r)$ parameters to the demand system. So provided $J > R$ we have added degrees of freedom.

To incorporate the information in (13) and (12) into the estimation algorithm rewrite the first order condition as $s + (p - \mathrm{mc})\Delta = 0$, where $\Delta_{i,j}$ is nonzero for elements of a row that are owned by the same firm as the row good. Then

$$p - \mathrm{mc} = \Delta^{-1}\sigma(\cdot).$$

Now substitute from (12) to obtain the cost disturbance as

$$\ln\bigl(p - \Delta^{-1}\sigma\bigr) - r'\theta^c = \omega(\theta), \tag{14}$$

and impose the restrictions that

$$Ef_j(w)\omega_j(\theta) = 0 \quad \text{at } \theta = \theta_0.$$

We add the empirical analogues of these moments to the demand side moments in (11) and proceed as in any method of moments estimation algorithm. This entails one additional computational step. Before we added the pricing system every time we evaluated a θ we had to simulate demand and do the contraction mapping for that θ. Now we also have to calculate the markups for that θ.

1.4.2. Adding micro data

There are a number of types of micro data that might be available. Sometimes we have surveys that match individual characteristics to a product chosen by the individual. Less frequently the survey also provides information on the consumer's second choice (see, for e.g. MicroBLP), or is a panel which follows multiple choices of the same consuming unit over time. Alternatively we might not have the original survey's individual choice data, but only summary statistics that provide information on the joint distribution of consumer and product characteristics [for a good example of this see Petrin's (2002) use of Consumer Expenditure Survey moments in his study of the benefits to the introduction of the minivan]. We should note that many of the micro data sets are choice based samples, and the empirical model should be built with this in mind [see, for e.g. MicroBLP (2004); for more on the literature on choice based sampling see Manski and Lerman (1977) and Imbens and Lancaster (1994)].

Since the model in (6) is a model of individual choice, it contains all the detail needed to incorporate the micro data into the estimation algorithm. Thus the probability of an individual with observed characteristics z_i choosing good j given (θ, δ) is given by

$$\Pr(j|z_i, \theta, \delta) = \int_v \frac{\exp[\delta_j + \Sigma_{kl} x_{jk} z_{il} \theta^{\mathrm{o}}_{kl} + \Sigma_{kl} x_{jk} \nu_{il} \theta^{\mathrm{u}}_{kl}]}{1 + \sum_q \exp[\delta_q + \Sigma_{kl} x_{qk} z_{il} \theta^{\mathrm{o}}_{kl} + \Sigma_{kl} x_{jk} \nu_{il} \theta^{\mathrm{u}}_{kl}]} f(\nu)\,\mathrm{d}(\nu). \tag{15}$$

1.4.2.1. What can be learned from micro data Assume temporarily that we can actually compute the probabilities in (15) analytically. Then we can use maximum likelihood to estimate $(\theta^{\mathrm{o}}, \theta^{\mathrm{u}})$. These estimates *do not* depend on any restrictions on the distribution of ξ. I.e. by estimating free δ_j coefficients, we are allowing for a free set of ξ_j.

On the other hand recall that

$$\delta_j = \Sigma_k x_{jk} \bar{\theta}_k + \xi_j.$$

So we cannot analyze many of the implications of the model (including own and cross price elasticities) without a further assumption which enables us to separate out the

effect of ξ from the effect of the x on δ (i.e. without the identifying assumption referred to above). The availability of micro data, then, *does not* solve the simultaneity problem. In particular, it does not enable us to separate out the effect of price from unobservable characteristics in determining aggregate demand. On the other hand there are a few implications of the model that can be analyzed from just the estimates of $(\delta, \theta^{\mathrm{o}}, \theta^{\mathrm{u}})$. In particular, estimates of consumer surplus from the products currently marketed (and hence "ideal" consumer price indices) depend only on these parameters, and hence do not require the additional identifying assumption.

Now say we wanted to use the data to estimate $\bar{\theta}$. In order to do so we need a further restriction so assume, as before, that we have instruments w, and can provide instrumental variable estimates of the $\bar{\theta}$. The number of observations for the instrumental variable regressions is the number of products. That is, at least if we chose to estimate $(\theta^{\mathrm{o}}, \theta^{\mathrm{u}})$ without imposing any constraints on the distribution of ξ, the precision of the estimates of $\bar{\theta}$ will depend only on the richness of the product level data. Moreover, IV regressions from a single cross-section of products in a given market are not likely to produce very precise results; in particular there is likely to be very little independent variance in prices. Since additional market level data is often widely available, this argues for integrating it with the micro data, and doing an integrated analysis of the two data sources.

One more conceptual point on survey data. What the survey data adds is information on the joint distribution of observed product and consumer attributes. We would expect this to be very helpful in estimating θ^{o}, the parameters that determine the interactions between z and x. There is a sense in which it also provides information on θ^{u}, but that information is likely to be much less precise. That is we can analyze the variance in purchases among individuals with the same choice set and the same value of z and use that, together with the i.i.d. structure of the ϵ, to try and sort out the variance-covariance of the ν. However this requires estimates of variances conditional on z, and in practice such estimates are often quite imprecise. This is another reason for augmenting cross-sectional survey data with aggregate data on multiple markets (or time periods) in an integrated estimation routine; then the observed variance in z could determine the θ^{o} and differences in choice sets could help sweep out the impact of the θ^{u} parameter.

When the data does have second choice information, or when we observe the same consuming unit purchasing more than one product, there is likely to be much more direct information on θ^{u}. This because the correlation between the *x-intensity* of the first choice and the second choice of a given individual is a function of both θ^{o} and the θ^{u} terms, and the θ^{o} terms should be able to be estimated from only the first choice data. A similar comment can be made for repeated choices, at least provided the utility function of the consuming unit does not change from choice to choice.

Table 2 illustrates some of these points. It is taken from MicroBLP where the data consisted of a single cross-sectional survey of households, and the market level data from the same year. The survey contained information on household income, the number of adults, the number of children, the age (of the head) of household, and whether their residence was rural, urban, or suburban (and all of these were used in the estimation).

Table 2
Price substitutes for selected vehicles, a comparison among models

Vehicle	Full model	Logit 1st	Logit 1st and 2nd	Sigma only
Metro	Tercel	Caravan	Ford FS PU	Civic
Cavalier	Escort	Caravan	Ford FS PU	Escort
Escort	Tempo	Caravan	Ford FS PU	Ranger
Corolla	Escort	Caravan	Ford FS PU	Civic
Sentra	Civic	Caravan	Ford FS PU	Civic
Accord	Camry	Caravan	Ford FS PU	Camry
Taurus	Accord	Caravan	Ford FS PU	Accord
Legend	Town Car	Caravan	Ford FS PU	LinTnc
Seville	Deville	Caravan	Ford FS PU	Deville
Lex LS400	MB 300	Econovan	Ford FS PU	Seville
Caravan	Voyager	Voyager	Voyager	Voyager
Quest	Aerostar	Caravan	Caravan	Aerostar
G Cherokee	Explorer	Caravan	Chv FS PU	Explorer
Trooper	Explorer	Caravan	Chv FS PU	Rodeo
GMC FS PU	Chv FS PU	Caravan	Chv FS PU	Chv FS PU
Toyota PU	Ranger	Caravan	Chv FS PU	Ranger
Econovan	Dodge Van	Caravan	Ford FS PU	Dodge Van

Source: From Berry, Levinsohn and Pakes (2004).

That study had particularly rich information on vehicle preferences, as each household reported its second as well as its first best choice.

Table 2 provides the best price substitutes for selected models from demand systems for automobiles that were estimated in four different ways: (i) the full model allows for both the z_i and the ν_i (i.e. for interactions between both observed and unobserved individual characteristics and product characteristics), (ii) the logit models that allow for only the z_i, and (iii) the σ's only model allows for only the ν_i. The most important point to note is that without allowing for the ν_i there is a clear IIA problem. The prevalence of the Caravan and the Full Size (FS) pickups when we use the logit estimates (the models without the ν_i) is a result of them being the vehicles with the largest market shares and the apparent absence of the observed factors which cause different households to prefer different product characteristics differentially. Comparing to column (iv) it is clear that the extent of preference heterogeneity caused by household attributes not in our data is large. MicroBLP also notes that when they tried to estimate the full model without the second choice information their estimates of the θ^{u} parameters were very imprecise; too imprecise to present. However when they added the second choice data they obtained both rather precise estimates of the contributions of the unobserved factors and substitution patterns that made quite a bit of sense. Finally we note that the fact that there was only a single year's worth of data made the estimates of $\bar{\theta}$ quite imprecise, and the paper uses other sources of information to estimate those parameters.

1.4.2.2. Computational and estimation issues: Micro data There are a number of choices to make here. At least in principal we could (i) estimate $(\theta^{o}, \theta^{u}, \delta)$ pointwise, or (ii) make an assumption on the distribution of ξ (e.g. $E[\xi|w] = 0$), and estimate $(\theta^{o}, \theta^{u}, \bar{\theta})$ instead of $(\theta^{o}, \theta^{u}, \delta)$. However the fact that ξ is a determinant of price, and price is in the x vector, makes it difficult to operationalize (ii). To do so it seems that one would have to make an assumption on the primitive distribution of ξ, solve out for equilibrium prices conditional on (θ, ξ, x), substitute that solution into the choice probabilities in (15), and then use simulation to integrate out the ξ and ν in the formula for those probabilities. This both involves additional assumptions and is extremely demanding computationally. The first procedure also has the advantage that its estimates of (θ^{o}, θ^{u}) are independent of the identifying restriction used to separate out the effect of ξ from the effect of x on $\bar{\theta}$.

Assume that we do estimate $(\theta^{o}, \theta^{u}, \delta)$. If there are a large number of products or J, this will be a large dimensional search (recall that there are J components of δ), and large dimensional searches are difficult computationally. One way to overcome this problem is to use the aggregate data to estimate δ conditional on θ from the contraction mapping in (9), and restrict the nonlinear search to searching for (θ^{o}, θ^{u}).

Finally since the probabilities in (15) are not analytic, either they, or some transform of them (like the score), will have to be simulated. There is now quite a bit of work on simulating the probabilities of a random coefficient logit model [see Train (2003) and the literature cited there]. Here we only want to remind the reader that in the applications we have in mind it is likely to be difficult to use the log (or a related) function of the simulated probabilities in the objective function. Recall that if $p^{ns}(\theta)$ is the simulated probability, and $p^{ns}(\theta) = p(\theta) + e^{ns}$, where e^{ns} is a zero mean simulation error, then

$$\log\bigl[p^{ns}(\theta)\bigr] \approx \log\bigl[p(\theta)\bigr] + \frac{e^{ns}}{p(\theta)} - \frac{(e^{ns})^2}{2 \times p(\theta)^2}.$$

So if the simulated probabilities are based on ns independent simulation draws each of which has variance $V(p(\theta))$ the *bias* in the estimate of the log probability will be approximately

$$E\log\bigl[p^{ns}(\theta)\bigr] - \log\bigl[p(\theta)\bigr] \approx -\frac{1}{2 \times ns \times p(\theta)},$$

and ns must be large relative to $p(\theta)$ for this bias to go away (this uses the fact that $\mathrm{Var}(p^{ns}(\theta)) \approx p(\theta)/ns$).

In many Industrial Organization problems the majority of the population do not purchase the good in a given period, and the probabilities of the inside goods are formed by distributing the remainder of the population among a very large number of goods. For example, in MicroBLP's auto example, only ten per cent of households purchase a car in the survey year, and that ten percent is distributed among more than two hundred models of cars. So it was common to have probabilities on the order of 10^{-4}. It should not be a surprise then that they chose to fit moments which were linear functions of the error in estimating the probabilities (they fit the covariances of car characteristics

and household characteristics predicted by the model to those in the data) rather than maximizing a simulated likelihood.

1.4.3. Identifying restrictions

Recall that the source of the endogeneity problem in the demand estimates is the correlation of the product specific unobservable, our ξ, with some of the observable characteristics of the product; in particular we are worried about a correlation of ξ with price. The contraction mapping in (9) is helpful in this respect as it delivers ξ as a linear function of observables. As a result, any of the standard ways of solving endogeneity problems in linear models can be employed here.

The most familiar way of dealing with endogeneity problems in linear models is to use instruments. The question then becomes what is an appropriate instrument for x's in the demand system, a question which has been discussed extensively in the context of perfectly competitive models of supply and demand. As in those models cost shifters that are excluded from demand and uncorrelated with the demand error are available as instruments. The familiar problem here is that input prices typically do not vary much; at least not within a single market. There are a couple of important exceptions. One is when production takes place in different locations even though the products are all sold in one market [as is common when investigating trade related issues, see Berry, Levinsohn and Pakes (1999)]. Another is when a subset of the x's are exogenous, the cost factors are differentially related to different x's, and the x-intensity of different product varies. In this case interactions between the cost factors and those x's should be useful instruments.

In addition to cost instruments, Nevo (2001) uses an idea from Hausman (1996) market-equilibrium version of the AIDS model, applied to a time-series/cross-section panel of geographically dispersed set of markets. The underlying assumption is that demand shocks are not correlated across markets while cost shocks are correlated across markets. The prices of goods in other markets then become instruments for the price of goods in a given market. Nevo (2001) studies breakfast cereals and so sources of common cost shocks include changes in input prices; sources of common demand shocks (which are ruled out) include national advertising campaigns.

In oligopoly markets prices typically sell at a markup over marginal cost. So if the product's own $(\tilde{x}_j, r_j)$'s are used as instruments, then so might the $(\tilde{x}_{-j}, r_{-j})$ of other products, giving us a lot of potential instruments. Moreover, if price setting models like the one in Equation (13) are appropriate (and recall that they often have a lot of explanatory power), the impact of the (x_{-j}, r_{-j}) on p_j will depend on whether the product's are owned by the same or by different firms. This type of reasoning dates back at least to Bresnahan (1987), who notes the empirical importance of the idea that markups will be lower in "crowded" parts of the product space and that they will be higher when "nearby" products are owned by the same firm. BLP and Berry, Levinsohn and Pakes (1999) rely on this sort of argument to propose the use of functions of rivals' observed product characteristics, and of the ownership structure of products, as instruments. Re-

latedly exogenous changes in competitive conditions across markets are also candidate instruments (say due to the size of the market, or heterogeneity in entry costs).

It is difficult to specify *a priori* how to make optimal use of the product characteristics to predict markups. Both BLP and Berry, Levinsohn and Pakes (1999) try approximations to the "optimal instrument" formula suggested by Chamberlain (1984). This assumes

$$E[\xi_j|\tilde{x}_j, \tilde{x}_{-j}, r_j, r_{-j}] = E[\omega_j|\tilde{x}_j, \tilde{x}_{-j}, r_j, r_{-j}] = 0,$$

homoscedasticity, and ignores the within market dependence induced by the market interactions. Chamberlain's results then imply that the optimal instrument for our problem is the derivative of these expectations with respect to the parameter vector.

In our context this will be a difficult to compute function of all the product characteristics. BLP tries to approximate this function "nonparametrically" using the exchangeable basis provided in Pakes (1994). Berry, Levinsohn and Pakes (1999), try an alternative approximation which is more direct, but also more computationally burdensome. They use a first-stage estimate of the parameter vector, θ, to recalculate equilibrium prices with all values of $\xi = \omega = 0$. They then compute the derivative of ξ and ω with respect to θ at the first stage estimate of θ and the new equilibrium prices, and use it as an instrument. I.e. instead of evaluating the mean of the derivative they evaluate the derivative at the mean of the disturbance vector. Note that the instrument is then a function only of exogenous variables, and so results in consistent estimators (even though they are not quite efficient).

So far we have assumed mean independence of the unobservable characteristics, and, as noted, there are plausible reasons to believe that product characteristics themselves are correlated with ξ. After all the product design team has at least some control over the level of ξ, and the costs and benefits of producing different levels of the unobservable characteristics might well vary with the observed characteristics of the product. One possible solution would be to completely model the choice of product characteristics, as in the dynamic models considered later in this chapter.

That said since p is typically not as hard to adjust as the other product characteristics, the relationship between ξ and $\tilde{x}$ does not seem to be nearly as direct as that between ξ and p (which is the reason it is often ignored; just as it was in traditional models of demand and supply). So one might be willing to make some reduced form assumption which allows us to proceed without all the detail of a dynamic game. In particular, one might try to use changes in demand over time, or across markets, for the same good to control for the influence of unobserved product characteristics.

For example, suppose that we observe demand for the same product over time. It might be reasonable to suppose that the product characteristics are correlated with the unobservable in the year of product introduction. However one might also argue that any changes in the level of unobserved characteristics over time are due to changes in either perception of the product or in customer service that have little to do with the initial x choices. So if t_0 were the date of introduction of the good we might assume that

$$\xi_{j,t} = \xi_{j,t_0} + \eta_{j,t+1}, \tag{16}$$

where $\eta_{j,t+1}$ is mean independent of the observed characteristics of all products. Alternatively we could assume that $\xi_{j,t}$ followed a first order Markov process with only ξ_{j,t_0}, and not the increments in the process, correlated with observed characteristics.

Relatedly if the data contains sales of the same product in many markets one could think of restrictions on how the unobservable for a single product changes across markets. The most straightforward example of this is to require ξ to be the same across markets. This is quite a powerful restriction, and one might question it on the basis of differences in the distribution of consumer preferences across markets that impact on their estimated ξ's. A weaker assumption would be that the difference between ξ's for the same product across markets is uncorrelated with the observed x. Similarly, some products within a market may differ only by the addition of some optional features and we could restrict the way that ξ changes across products that vary only in their options.

1.5. Problems with the framework

We have motivated our discussion on demand estimation by noting how the recent literature dealt with the problems that arose in using representative agent models in product space. There are many senses, however, in which the framework outlined above can be too restrictive for particular problems. This section reviews some of the more obvious of them. The impact of these problems depend upon the market one is analyzing and the issues one is focusing on. Also, at least partial solutions to some of these problems are available, and we will direct the reader to them where we can. In large part, however, this section is an outline of agendas for future research on demand estimation for I.O. problems.

We begin with multiple choice and/or dynamics, and then come back to the problem in the static discrete choice model considered above. Most empirical studies simply ignore issues related to multiple choices and/or dynamics. The hope is that the estimated demand system is still the best currently available approximation for analyzing the question of interest. To us the surprising part of the results of those studies is that the framework seems to provide a "reasonable" approximation to substitution patterns, and even more surprisingly, a reasonable approximation to pricing patterns. This despite the fact that we know that consumers' demands and the market equilibrium outcomes are products of much more complicated processes than those we model. Even so, as will become clear presently, there are a number of issues of importance to I.O. which cannot be studied empirically without a more detailed understanding of multiple choice and/or the dynamic aspects of demand.

Multiple units of demand

There are many situations for which a model based on the choice of either one or zero units of a good does not match reality.[11] Models for choosing a finite number of units

[11] Dubin and McFadden (1984) provide an earlier example with one discrete choice and one continuous choice.

from a set of substitute goods require a specification for the utility from multiple units. Then, at least in principle, we are back to a discrete choice for "tuples" of goods. However to maintain tractability when the number of units can grow large the specification is likely to require constraints which cut down the choice set by implying that some choices are dominated by others (otherwise the size of the choice set grows as J^C, where J is the number of products and C is the maximum number of purchases).

One example of the use of such constraints is Hendel's (1999) two-stage multiple-unit/multiple good framework for the demand of a firm for computers. He simplifies the problem by imagining that the firm faces a random, discrete number of tasks. For each task, it chooses only one type (brand) of computer and, according to the random size of the tasks, a number of computers to purchase. This explicitly accounts for decisions to purchase multiple units of multiple kinds of goods.

Gentzkow (2004) considers a problem with a small number of goods, but where there are a small number of choices. In that study of online and print newspapers, some of the goods are potentially complements, and this requires a different set of modifications. Moreover, as Gentzkow shows the determination of whether goods are in fact complements or substitutes interacts with the issue of the form of consumer heterogeneity in subtle ways reminiscent of the initial condition problems in panel data estimation [see Heckman (1981)].

A related problem involves continuous choice over multiple goods. If all goods are purchased in some positive amount by every consumer, then a traditional continuous demand approach, equating marginal rates of substitution across all goods, is appropriate. But many real-world consumer data problems involve a large number of goods with many zero purchase decisions and many positive purchase decisions. Chan (2002) considers the Kuhn–Tucker version of the traditional continuous choice problem to study soft drink purchases.

Dynamic demand

Yet another set of problems arises when the demand for the good is inherently dynamic, as occurs with either durable, storable, or experience goods. Models which are appropriate for dynamic demand estimation can become quite complex; they require forward looking consumers whose behavior depends on the likely distribution of future (as well as current) offerings. Moreover, in a complete model these future offerings would, in turn, depend on producer's perceptions of consumer demand. A number of new studies make simplifying assumptions which allow them to make some headway.

Both Hendel and Nevo (2002) and Erdem, Imai and Keane (2003) consider a problem of durable good demand in an explicitly dynamic framework. They consider shopping decisions when consumers are allowed to store purchases, and use a reduced form assumption on the process generating prices. It has been clear to I.O. economists for some time that we are going to have to model intertemporal substitution of this form in order to understand "sales" in retail markets [see Sobel (1984)].

Two problems in this kind of study are that the rate of consumption (inventory reduction) at home is typically not observed and the dimension of the state space (which involves both the current price vector, which predicts future prices, and also the vector of household inventories of different brands). In these models new purchases are added to a single-index of home inventories, with different brands of product receiving different utility weights in the inventory stock. This single index of inventories reduces the dimensionality of the state space. Another simplifying assumption is that unobserved household consumption follows a simple rule.

Esteban and Shum (2002) consider a model of durable automobile purchases. They assume a used-car market with zero transaction costs. The zero transaction costs imply that the joint distribution of past choices and consumer characteristics are not a state variable of the problem. Under these assumptions they are able to derive empirical implications about the dynamic pricing problem of the durable goods manufacturer (in determining current price the manufacturer has to worry about future aggregate supply of the used goods). Many, if not most, manufacturing goods are durable.

Studies of demand for advertised experience goods include Erdem and Keane (1996), Ackerberg (2003), and Crawford and Shum (2007). All of these papers feature Bayesian consumers who learn both from experience and from adverting. This leads to a fairly complex dynamic programming problems for the consumer. The studies largely ignore the firm's endogenous pricing and advertising decisions.

Problems with the static discrete choice specification

There are also aspects of the static discrete choice specification of the model outlined above whose flexibility, and/or implications, are not yet well understood. One such issue is whether the second derivatives of the demand function are very flexibly estimated. This will determine whether two goods are strategic substitutes or strategic complements, and hence has implications for the analysis of the structure of strategic interaction, and appears to be largely unexplored in the current literature. More generally there are a host of questions on what we can learn nonparametrically about the structure of demand from different kinds of data that we have not touched on here (for a discussion of some of them, see Matzkin's contribution to this volume).

A second such issue concerns the role of the i.i.d. "idiosyncratic match values", the ϵ_{ij}'s, in the models above. These are added to the model largely for computational convenience; they do not seem to match any omitted causal demand determinant. Moreover, the presence of the ϵ_{ij} has implications. They imply that each product is "born" with a distribution of consumer tastes whose conditional distribution, conditional on the tastes for other products, has support that ranges from minus to plus infinity. This implies that every conceivable product, no matter its characteristics and price, will have a strictly positive (though perhaps quite tiny) expected market share.

Given the standard ϵ_{ij}'s, each product will also have a positive cross-price effect with every other product: competition is never completely local. Perhaps most problematic, it also implies that if we define a consumer by a (z, ν) combination, every consumer's

utility will grow without bound as we increase the number of products – *regardless* of the characteristics or prices of the new products that are introduced. As a result there is a worry about the ability of the model in (6) to provide an adequate approximation to the benefits from introducing new goods.[12]

To investigate these issues more fully, Berry and Pakes (2005) consider a "pure characteristic" model of demand. That model is exactly the model in Equation (6) once we omit the ϵ_{ij} terms. They consider the analytic properties of the model, then provide an estimation algorithm for it and explore its computational properties, and finally provide Monte Carlo evidence on its performance. Song (2004) has used this model to evaluate the gains from new semiconductor chips. The pure characteristics model is somewhat more computationally burdensome then the model in Equation (6), largely because the equation for solving for δ for that model (the analogue to Equation (9)) is not necessarily a contraction with modulus less than one. On the other hand its shares are easier to simulate to sufficient accuracy. However the jury is still out on the major question; the question of whether the pure characteristic model tends to provide a better approximation to the consumer surplus gains from new goods then the model with the ϵ_{ij}.

Berry and Pakes (2005) and Bajari and Benkard (2005) discuss two different versions of the "pure characteristics" model with "no ϵ"s. Berry and Pakes (2005) consider a discrete choice version of the model, with a utility function of

$$u_{ij} = x_j \beta_i - \alpha_i p_j + \xi_j, \tag{17}$$

where β_i and α_i are random coefficients associated with consumer i's tastes for characteristics and price of product j. Berry and Pakes suggest a BLP-style estimation algorithm.

In contrast, Bajari and Benkard (2005) obtain an estimate of the unobservable demand component, ξ_j, from the pricing side of the model rather than the demand side. The argument is that in a "pure characteristics" model, prices must be strictly increasing in ξ conditional on other x's. Following on recent econometric literature, they show that a monotonic transformation of the ξ can be obtained from data on prices and x's. This transformed ξ is then used in the demand-side analysis to control for unobserved characteristics. Note, however, that consistency of this approach relies on asymptotics in the number of products, and further requires the assumption that products enter the market in such a way that eventually they "fill up" the product space (i.e., for every product, it is assumed that eventually there will be other products whose observed characteristics are arbitrarily close to those of the given product). In practice it is clear that the approach

[12] We hasten to note that estimating the consumer surplus generated by new products is an extremely difficult task in any framework. This is because we typically do not have data on the demand for new products at prices that are high enough to enable us to estimate the reservation prices of a large fraction of consumers. The characteristic based demand model does use slightly more information in its estimation of consumer surplus gains than do demand models in product space, since it uses the price variance for products with similar characteristics. However the results are still not terribly robust. Petrin (2002), for example, reports large differences in consumer surplus gains from differences in specifications and data sources.

requires data with many products per market, but there has not been enough experience to date to know what "many" means in this context.

1.6. Econometric details

This subsection summarizes results from Berry, Linton and Pakes (2004) who provide limit theorems for the parameter estimates from differentiated product models. The actual form of the limit distributions depends on the type of data and type of model. We will focus on the case where only one cross section of market level data is available. Our purpose is to give the reader some indication of how the various estimation errors that have been introduced are likely to effect the parameter estimates, and this is the simplest environment in which to show that.[13]

Recall that the objective function minimized in the estimation algorithm or Equation (11) is a norm of

$$G_J\left(\theta, s^n, P^{ns}\right) = \frac{1}{J}\sum_{j=1}^{J} \xi_j\left(\theta, s^n, P^{ns}\right) f_j(w).$$

The ξ_j are defined implicitly as the solution to the system

$$s_j^n = \sigma_j\left(\xi, x; \theta, P^{ns}\right),$$

where $\sigma(\cdot)$ is defined in (8), the w satisfy $E[\xi|w, \theta_0] = 0$, s^n is the observed vector of market shares, and P^{ns} is notation for the vector of simulation draws used to compute the market shares predicted by the model.

The objective function, $\|G_J(\theta, s^n, P^{ns})\|$, has a distribution determined by three independent sources of randomness: randomness generated from the draws on the product characteristics (both observed and unobserved, in the full model these are vectors $\{\xi, \tilde{x}, r, \omega\}$), randomness generated from the sampling distribution of s^n, and that generated from the simulated distribution P^{ns}. Analogously there are three dimensions in which our sample can grow: as n, as ns, and as J grow large.

The limit theorems allow different rates of growth for each dimension. Throughout we take pathwise limits, i.e. we write $n(J)$ and $ns(J)$, let $J \to \infty$, and note that our assumptions imply $n(J), ns(J) \to \infty$ at some specified rate. Note also that both s^n and $\sigma(\xi, \theta, P)$ take values in R^J, where J is one of the dimensions that we let grow in our limiting arguments. This is an unusual feature of the econometric model and causes complications in the limiting arguments. As will become obvious sampling error (error

[13] Cases in which there is data from many regional markets but the same goods are sold in each of them will still have to deal with limits as the number of products grows large; it is just that then we might also want to let the number of markets increase as we increase the number of products. Also in cases with regional markets the computational problems we highlight will be even more severe, as then we will have to compute ξ separately in each different market.

in s^n) plays an analogous role to simulation error (error in P^{ns}), so for notational simplicity assume that n is sufficiently large that we do not need to worry about sampling error. When there is no sampling (simulation) error we set n (ns) equal to zero.

We need to find an approximation for the objective function which allows us to separate out the roles of the three sources of error. To this end write

$$\xi\left(\theta, s^0, P^{ns}\right) = \xi\left(\theta, s^0, P^0\right) + \left\{\xi\left(\theta, s^0, P^{ns}\right) - \xi\left(\theta, s^0, P^0\right)\right\}. \tag{18}$$

The function $\sigma(\xi, \theta, P)$ is differentiable in ξ, and its derivative has an inverse, say

$$H^{-1}(\xi, \theta, P) = \left\{\frac{\partial\sigma(\xi, \theta, P)}{\partial\xi'}\right\}^{-1}.$$

Abbreviate $\sigma_o(\theta, s, P) = \sigma(\xi(s, \theta, P), \theta, P)$ and $H_o(\theta, s, P) = H(\xi(s, \theta, P), \theta, P)$, and let

$$\sigma\left(\xi, P^{ns}, \theta\right) = \sigma\left(\xi, P^0, \theta\right) + \varepsilon^{ns}(\theta).$$

Then from the fact that we obtain ξ from $\sigma(\cdot) = \sigma(\xi, P^0, \theta) + \varepsilon^{ns}(\theta)$ it follows that

$$\xi\left(\theta, s^0, P^{ns}\right) = \xi\left(\theta, s^0, P^0\right) + H_o^{-1}\left(\theta, s^0, P^0\right)\left\{\varepsilon^{ns}(\theta)\right\} + r\left(\theta, s^n, P^{ns}\right),$$

where $r(\theta, s^n, P^{ns})$ is a remainder term. Substituting into (18)

$$G_J\left(\theta, s^n, P^{ns}\right) = G_J\left(\theta, s^0, P^0\right) + \frac{1}{J}z' H_o^{-1}\left(\theta, s^0, P^0\right)\left\{-\varepsilon^{ns}(\theta)\right\} + \frac{1}{J}z' r\left(\theta, s^n, P^{ns}\right).$$

The limit theorems in Berry, Linton and Pakes (2004) work from this representation of $G_J(\theta, s^n, P^{ns})$. To prove consistency they provide conditions which insure that: (i) the second and third terms in this equation converge to zero in probability uniformly in θ, and (ii) an estimator which minimized $\|G_J(\theta, s^0, P^0)\|$ over $\theta \in \Theta$ would lead to a consistent estimator of θ^0.

Asymptotic normality requires, in addition, local regularity conditions of standard form, and a limiting distribution for $H_o^{-1}(\theta, s^0, P^0)\{-\varepsilon^{ns}(\theta)\}$. The rate needed for this limit distribution depends on how the elements of the $J \times J$ matrix $H_o^{-1}(\theta, s^0, P^0)$ grow, as J gets large. It is easiest to illustrate the issues that can arise here by going back to the simple logit model.

In that model: $u_{i,j} = \delta_j + \epsilon_{i,j}$, with the $\{\epsilon_{i,j}\}$ distributed i.i.d. type II extreme value, and $\delta_j = x_j\bar{\theta} + \xi_j$. Familiar arguments show that $\sigma_j = \exp[\delta_j]/(1 + \sum_q \exp[\delta_q])$, while $\sigma_0 = 1/(1 + \sum_q \exp[\delta_q])$. In this case the solution to the contraction mapping in (9) is analytic and

$$\xi_j\left(\theta, s^o, P^o\right) = \left(\ln\left[s_j^o\right] - \ln\left[s_0^o\right]\right) - x_j\beta.$$

Thus in this simple case

$$\left.\frac{\partial\xi}{\partial s_j}\right|_{s^o} = \frac{1}{s_j^o}.$$

Now consider how randomness affects the estimate of $\xi_j(\theta)$. In the simple logit model the only source of randomness is in the sampling distribution of s^n. That is we observe the purchases of only a finite random sample of consumers. Letting their shares be s^n we have, $s^n - s^o = \epsilon^n$. The first order impact of this randomness on the value of our objective function at any θ will be given by

$$H_o^{-1}(\theta, s^0) \times \epsilon_n = \left.\frac{\partial \xi}{\partial s}\right|_{s=s^0} \times \epsilon^n.$$

This contains expressions like $\epsilon_j^n \frac{1}{s_j^o}$. In the logit model as $J \to \infty$, $s_j^o \to 0$. So as J grows large the impact of any given sampling error grows without bound.

A similar argument holds for the estimator of BLP's model, only in this more complicated model there are two sources or randomness whose impacts increase as J grows large, sampling error and simulation error. Consequently Berry, Linton and Pakes show that to obtain an asymptotically normal estimator of the parameter vector from this model both n and ns must grow at rate J^2. Note the similarity here to the reason that simulation error is likely to make use of maximum likelihood techniques with survey data computationally demanding; i.e. the impact of the simulation error on the objective function increases as the actual shares get smaller. The computational implication here is that for data sets with large J one will have to use many simulation draws, and large samples of purchasers, before one can expect to obtain an accurate estimator whose distribution is approximated well by a normal with finite variance.

Interestingly, this is not the case for the pure characteristic model discussed in the last subsection. We will not provide the argument here but Berry, Linton and Pakes (2004) show that in that model both n and ns need only grow at rate J (and depending on the pricing equilibrium, sometimes slower rates will do), for the normal limit distribution to be appropriate. This gives the pure characteristic model a computational advantage in calculating shares, though, as noted above, it is harder to compute the analogue of the contraction mapping in (9) for the pure characteristics model, so it can still be computationally demanding.

1.7. Concluding remark

The last decade has seen a rather dramatic change in the way I.O. researchers analyze demand systems. There now is a reasonably substantial body of academic research using the new techniques, and it seems to indicate that, at least for many situations, they allow us to get better approximations to substitution patterns and the likely demand for new goods than had been possible previously. Perhaps not surprisingly then, the techniques have been picked up, to varying extents, by: the consulting community, various government offices, and even by a part of the business community. On the other hand, as we have tried to emphasize, there are empirically important issues and data sets that the new techniques are not able to analyze – at least not without substantial further developments. We welcome those developments. Moreover, we hope that they will not

be judged by any absolute criteria but rather by the simple test of whether they allow for improvements in our ability to empirically analyze one or more issue of substantive interest.

2. Production functions

As noted in the introduction, the advent of new micro data sets on the inputs and outputs from the production process has generated a renewed interest in the estimation of production functions and their use in the analysis of productivity. We begin this section by reviewing the basic simultaneity and selection issues that the recent literature on production function estimation has faced. We then consider the traditional solutions to these issues, pointing out why those solutions are not likely to be terribly helpful in our context.

Next we introduce an approach based on explicit models of input choices and exit decisions that was first introduced in a paper by Olley and Pakes (1996). Our presentation of the Olley–Pakes model will stress the assumptions they used which either we, or others before us, see as questionable (at least in certain environments). These include assumptions on: the timing of input choices, the cost of changing the levels of different inputs over time, the process by which productivity evolves over time, and the relationship of investment to that process. The rest of the section focuses on ways of testing these assumptions, and details recently proposed modifications to the estimation procedure which might be used when they seem appropriate.

2.1. Basic econometric endogeneity issues

We can illustrate all issues that will concern us with simple Cobb–Douglas production technology

$$Y_j = A_j K_j^{\beta_k} L_j^{\beta_l}$$

with one output (Y_j) and two inputs; capital (K_j) and labor (L_j). A_j represents the Hicksian neutral efficiency level of firm j, which is unobserved by the econometrician.[14]

Taking natural logs results in a linear equation

$$y_j = \beta_0 + \beta_k k_j + \beta_l l_j + \epsilon_j, \tag{19}$$

where lowercase symbols represent natural logs of variables and $\ln(A_j) = \beta_0 + \epsilon_j$. The constant term β_0 can be interpreted as the mean efficiency level across firms, while

[14] The methods discussed in this chapter are equally applicable to many other production functions. As we shall see the major requirements will be that variable inputs have positive cross-partials with productivity, and that the value of the firm is increasing in fixed inputs.

ϵ_j is the deviation from that mean for firm j. ϵ_j might represent innate technology or management differences between firms, measurement errors in output, or unobserved sources of variance in output caused by weather, machine breakdowns, labor problems, etc.

We have known since Marschak and Andrews (1944) that direct OLS estimation of (19) is problematic. The problem is that the right-hand side variables, capital and labor, are generally chosen by the firm. If the firm has knowledge of its ϵ_j (or some part of ϵ_j) when making these input choices, the choices will likely be correlated with ϵ_j. For example, suppose that firms operate in perfectly competitive input and output markets (w_j, r_j, and p_j being the prices of labor, capital, and output, respectively), that capital is a fixed input, that firms perfectly observe ϵ_j before choosing labor, and that firms' current choices of labor only impact current profits and have no effect on future profits. Then the firm's optimal short-run choice of labor input is given by

$$L_j = \left[\frac{p_j}{w_j}\beta_l e^{\beta_0+\epsilon_j} K_j^{\beta_k}\right]^{\frac{1}{1-\beta_l}}. \tag{20}$$

Since choice of L_j (and thus l_j) depends directly on ϵ_j, OLS will generate biased coefficient estimates. In more general models, firms' choices of K_j will also typically be correlated with ϵ_j.[15]

There is a second, less well documented, endogeneity problem often inherent in OLS estimation of (19). Firm level datasets usually have a considerable level of attrition. For example, over a wide range of manufacturing industries, Dunne, Roberts and Samuelson (1988) find exit rates higher than 30% between 5 year census pairs. In applied work, one only has data on firms prior to exiting. If firms have some knowledge of ϵ_j prior to exiting, the firms that continue to produce will have ϵ_j draws from a selected sample, and the selection criteria will be partially determined by the other fixed inputs. Again as a simple example, suppose that firms are monopolies that are exogenously endowed with different fixed levels of capital. Firms then observe ϵ_j, decide whether to exit or not, and choose labor and produce if they have not exited. Also for simplicity suppose that after production firms disappear, so that the firms have no dynamic considerations. Firms in this situation will have an exit rule of the following form:

$$\chi(\epsilon_j, K_j; p_j, w_j, \beta) = 0 \text{ (or exit)} \quad \text{iff} \quad \Pi(\epsilon_j, K_j; p_j, w_j, \beta) < \Psi,$$

where β is the set of parameters (β_0, β_l, β_k) and Ψ is the nonnegative selloff value of the firm. Π is the argmax (over the variable input labor) of variable profits. This condition states that firms exit if variable profits are not at least as high as the selloff value of the firm.[16]

[15] Empirical results have lead practitioners to conclude that most often the bias imparted on the labor coefficient β_l is larger than the bias imparted on the capital coefficient β_k. This is consistent with models of input choice where labor is more easily adjustable than capital (i.e. labor is a "more variable" input than capital). The intuition here is that because it is more quickly adjustable, labor is more highly correlated with ϵ_j.

[16] This is a very simple example of an exit rule. More realistic models of exit would be dynamic in nature and distinguish between fixed and sunk costs; see the discussion below.

The key point is that this exit condition will generate correlation between ϵ_j and K_j *conditional* on being in the dataset (i.e. on not exiting). In the Cobb–Douglas case, both ϵ_j and K_j positively impact variable profits. As a result, selection will generate *negative* correlation between ϵ_j and K_j, since firms with higher K_j will be able to withstand lower ϵ_j without exiting. Thus, even if K_j is exogenous in the sense that it is uncorrelated with ϵ_j in the entire population of *potentially* active firms, selection can generate negative correlation in one's sample.

2.2. Traditional solutions

As is often the case, the two traditional solutions to these endogeneity problems are instrumental variables and fixed effects. Before discussing these approaches, we make two slight changes to our basic model. First, to explicitly consider the use of longitudinal panel data, we index our variables by time t. Second, to be precise about where exactly the endogeneity problems are coming from, we divide the unobservable ϵ_{jt} into two components, ω_{jt} and η_{jt}, i.e.

$$y_{jt} = \beta_0 + \beta_k k_{jt} + \beta_l l_{jt} + \omega_{jt} + \eta_{jt}. \tag{21}$$

The η_{jt} here are intended to represent unobservables that are *not* observed (or predictable) by the firm before input and exit decisions at time t. As such, they will not be correlated with these choices of inputs or exit behavior. On the other hand we do allow the possibility that ω_{jt} is observed (or predictable) by firms when they choose inputs and make exit decisions. Intuitively, ω_{jt} might represent factors like managerial ability at a firm, expected down-time due to machine breakdowns or strikes, or the expected rainfall at a farm's location. η_{jt} might represent deviations from expected breakdown rates in a particular year or deviations from expected rainfall at a farm. Another valid interpretation of η_{jt} is that it is classical measurement error in y_{jt} that is uncorrelated with inputs and exit decisions. The basic point here is that we have consolidated our endogeneity problems into ω_{jt}. η_{jt} is not a concern in that regard. We will often refer to ω_{jt} as the firms "unobserved productivity".

2.2.1. Instrumental variables

Instrumental variables approaches rely on finding appropriate instruments – variables that are correlated with the endogenous explanatory variables but do not enter the production function and are uncorrelated with the production function residuals. Fortunately, the economics of production suggests some natural instruments. Examining input demand functions (such as (20)) suggests that input prices (r_{jt} and w_{jt}) directly influence choices of inputs. In addition, these prices do not directly enter the production function. The last necessary condition is that the input prices need to be uncorrelated with ω_{jt}. Whether this is the case depends on the competitive nature of the input markets that the firm is operating in. If input markets are perfectly competitive, then input prices should be uncorrelated with ω_{jt} since the firm has no impact on market prices.

This is the primary assumption necessary to validate input price instruments. Note why things break down when firms have market power in input markets. If this is the case, input prices will be a function of the quantity of purchased inputs, which will generally depend on ω_{jt}.[17]

While using input prices as instruments may make sense theoretically, the IV approach has not been uniformly successful in practice. We believe there are at least four reasons for this. First input prices are often not reported by firms, and when firms do report the labor cost variable, i.e. w_{jt}, is often reported in a way that makes it difficult to use. Labor costs are typically reported as average wage per worker (or average wage per hour of labor). Optimally, we would want this variable to measure differences in exogenous labor market conditions faced by firms. Unfortunately, it may also pick up some component of unmeasured worker quality. Suppose we as econometricians do not observe worker quality, and that some firms employ higher quality workers than others. Presumably, the firms with higher quality workers must pay higher average wages. The problem here is that unobserved worker quality will enter the production function through the unobservable ω_{jt}. As a result, ω_{jt} will likely be positively correlated with observed wages w_{jt}, invalidating use of w_{jt} as an instrument.

Second, to use prices such as r_{jt} and w_{jt} as instruments requires econometrically helpful variation in these variables. While input prices clearly change over time, such time variation is not helpful when one wants to allow flexible effects of time in the production function (e.g. allowing β_0 to be a flexible function of t). One generally needs significant variation in r_{jt} and w_{jt} *across firms* to identify production function coefficients. This can be a problem as we often tend to think of input markets as being fairly national in scope. One might not expect, for example, the price of capital or labor market conditions to vary that much between states. Summarizing, to use the IV approach one: (1) has to observe significant variation in input prices across firms in the data, and (2) believe that this variation is due primarily to differences in exogenous input market conditions, *not* due to differences in unobserved input quality.

A third problem with IV is that it relies fairly strongly on an assumption that ω_{jt} evolves exogenously over time, i.e. firms do not choose an input that affects the evolution of ω_{jt}. Allowing ω_{jt} to be affected by *chosen* inputs that we do not control for is very problematic econometrically for the IV approach, for then it would be hard to imagine finding valid instruments for observed input choices. One would need to find variables that affect one input choice but that do not affect other input choices. In general this will be hard to do, since individual input choices typically depend on all input prices.

[17] Another possible instrument is output prices, as long as the firm operates in competitive output markets. These instruments have been used less frequently, presumably because input markets are thought to be more likely to be competitive. Other related instruments are variables that shift either the demand for output or the supply of inputs. While these types of instruments are typically harder to come by, one can argue that they are valid regardless of the competitive nature of input or output markets.

Finally, the IV approach only addresses endogeneity of input choice, not endogenous exit. Endogenous exit will tend to invalidate the direct use of input prices as instruments. The reason for this is that it is probable that the exit decision will be based in part on input prices. For example, we might expect that firms who face higher input prices to be more likely to exit (i.e. would exit at a higher ω_{jt}). This is likely to generate positive correlation between the instruments and the residuals in the production function. While direct application of IV in this situation is problematic, it is possible that one could combine the population orthogonality assumptions with a selection model [e.g. Gronau (1974), Heckman (1974, 1976, 1979)] to generate a consistent estimator of the production function parameters.

2.2.2. *Fixed effects*

A second traditional approach to dealing with production function endogeneity issues is fixed effects estimation. In fact, fixed effects estimators were introduced to economics in the production function context [Hoch (1962), Mundlak (1961)]. Fixed effects approaches make explicit use of firm panel data. The basic assumption behind fixed effects estimation is that unobserved productivity ω_{jt} is *constant* over time, i.e.

$$y_{jt} = \beta_0 + \beta_k k_{jt} + \beta_l l_{jt} + \omega_j + \eta_{jt}. \tag{22}$$

This allows one to consistently estimate production function parameters using either mean differencing, first differencing, or least squares dummy variables estimation techniques. First differencing, for example, leads to

$$y_{jt} - y_{jt-1} = \beta_k(k_{jt} - k_{jt-1}) + \beta_l(l_{jt} - l_{jt-1}) + (\eta_{jt} - \eta_{jt-1}). \tag{23}$$

Given the assumption that the η_{jt}'s are uncorrelated with input choices $\forall t$,[18] this equation can be consistently estimated by OLS.[19] Note that this approach simultaneously solves the selection problem of endogenous exit, at least if exit decisions are determined by the time invariant ω_j (and not by the η_{jt}'s). While fixed effects approaches are fairly straightforward and have certainly been used in practice, they have not been judged to be all that successful at solving endogeneity problems in production functions either. Again, there are a number of reasons why this may be the case.

First, it is clearly a strong assumption that ω_j is constant over time. This is especially true given the longer time frames for which panel data is now becoming available. In addition, researchers are often interested in studying periods of data containing major economic environmental changes (e.g. deregulation, privatization, trade policy

[18] The assumption that η_{jt}'s are uncorrelated with input choices (and possibly entry/exit decisions) at all time periods t is often described as a "strict" exogeneity assumption. One can often estimate these fixed effects models under weaker, "sequential" exogeneity assumptions, i.e. that η_{jt}'s are uncorrelated with input choices at all time periods $\leqslant t$. See Wooldridge (2002) for a discussion of these issues.

[19] Note that generic OLS standard errors are wrong because the residuals will be correlated across observations.

changes, ...). Typically these changes affect different firms' productivities differently, and those firms that the change impacts positively will be more likely to increase their inputs and less likely to exit.[20]

A second potential problem with fixed effects estimators is that when there is measurement error in inputs, fixed effects can actually generate worse estimates than standard level (OLS) estimators. Griliches and Hausman (1986) note that when inputs are more serially correlated over time than is input measurement error, differencing can lower the signal to noise ratio in the explanatory variables.[21] This can generate higher biases in fixed effects estimators than in OLS estimators, even if ω_j is constant over time and correlated with the explanatory variables.[22]

Lastly, fixed effects estimators simply have not performed well in practice. One often gets unreasonably low estimates of capital coefficients.[23] Even one of the seminal papers, Hoch (1962), for example, finds estimates of returns to scale around 0.6 – almost certainly an unrealistically low number. Another empirical finding that appears to contradict the fixed effect assumption concerns the comparison of fixed effects estimates on balanced panels (containing only observations for firms appearing throughout the sample) to those on the full panel. As mentioned above, if ω_j is constant over time, fixed effects estimation completely addresses selection and input endogeneity problems. As a result, one should obtain similar fixed effects estimates whether one uses the balanced sample or the full sample. Olley and Pakes (1996), for example, find very large differences in these two estimates, suggesting that the fixed effects assumption is invalid. That said, whether or not one takes fixed effects estimates as serious estimates of structural production function parameters, the fixed effect decomposition of variation into within and between components often provides a useful reduced form look at a dataset.

2.3. *The Olley and Pakes (1996) approach*

A recent paper by Olley and Pakes (1996) (henceforth OP) takes a different approach to solving both the simultaneity and selection problems inherent in production function estimation. Their empirical context is that of telecommunications equipment producers

[20] The restriction that ω_j is constant over time is one that has been relaxed (in parametric ways) in the dynamic panel data literature, e.g. Chamberlain (1984), Arellano and Bond (1991), Arellano and Bover (1995), and Blundell and Bond (1999). For example, these methods can allow ω_{jt} to be composed of a fixed effect plus an AR(1) process.

[21] By signal to noise ratio, Griliches and Hausman mean the variance in an observed explanatory variable due to true variance in the variable, vs. variance in the observed explanatory variable due to measurement error. This signal to noise ratio is inversely related to the bias induced by measurement error.

[22] Note that in this case (i.e. when there is measurement error in inputs), both fixed effects and OLS estimators are biased. Also, note that the more structural approaches discussed later in this chapter are likely also prone to this critique.

[23] "Unreasonable" is clearly not a completely precise statement here. We are referring to cases where the estimated capital coefficient is considerably below capital's cost share or where returns to scale are extremely low.

(using data from the US Census Bureau's longitudinal research database). The basic empirical goal is to measure the impact of deregulation and the breakup of AT&T on measures of plant level productivity. Our focus is on the OP methodology for addressing the endogeneity problems rather than the actual empirical results.

As we work through the OP approach, it is useful to keep in mind three types of assumptions that will be important in the approach. First there are assumptions on timing and the dynamic nature of inputs. Timing refers to the point in time when inputs are chosen by the firm relative to when they are utilized in production. "Dynamic nature" refers to whether the input choices of the current period affect the cost of input use in future periods; if it does not the input is labelled nondynamic and if it does the input is labelled as dynamic (and its current value becomes a "state variable" in the problem). Second, there will be a *scalar unobservable* assumption. This assumption limits the dimensionality of the econometric unobservables that impact firm behavior. Third, there will be a *strict monotonicity* assumption on the investment demand function – basically that investment level is strictly monotonic in the scalar unobservable (at least for firms whose investment level is strictly positive). We will see that this last assumption can be generated by more basic assumptions on economic primitives. While some of these assumptions can be relaxed in various ways, we delay that discussion until the next subsection.

Lastly, note that we focus on how to use the OP methodology in practice. We do not address the higher level technical aspects of the methodology, e.g. semiparametric consistency proofs and alternative standard error derivations for their two-step estimators. For discussion of these issues, e.g. see Pakes and Olley (1995) and the literature they cite. One might also look at Wooldridge (2004), who presents a concise, one-step, formulation of the OP approach for which standard error derivations are more straightforward.[24] This one-step approach may also be more efficient than the standard OP methodology.

The rest of this section discusses in detail the workings of the OP methodology. We start by describing a simple, bare bones, version of the model and methodology that ignores potential selection problems. We then move on to the full OP model, which does address selection. Lastly, we discuss caveats and extensions of the OP procedure.

2.3.1. The model

The OP approach considers firms operating through discrete time, making production choices to maximize the present discounted value (PDV) of current and future profits. The environment is as follows. First, the assumed production function is similar to (21), with an additional input a_{jt}

$$y_{jt} = \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + \beta_l l_{jt} + \omega_{jt} + \eta_{jt} \tag{24}$$

[24] Though Woolridge deals with input endogeneity, he does not explicitly consider the selection issue. However similar ideas can be used when one needs to incorporate selection corrections.

the natural log of the age (in years) of a plant. The interest in the age coefficient stems from a desire to separate out cohort from selection effects in determining the impact of age of plant on productivity.

Second, unobserved productivity ω_{jt} is assumed to follow an exogenous first order Markov process. Formally,

$$p\left(\omega_{jt+1}\middle|\{\omega_{j\tau}\}_{\tau=0}^{t}, I_{jt}\right) = p(\omega_{jt+1}|\omega_{jt}), \tag{25}$$

where I_{jt} is the firm's entire information set at time t. This is simultaneously an *econometric* assumption on unobservables *and* an *economic* assumption on how firms form their perceptions on (i.e. learn about) the evolution of their productivity over time. Specifically, a firm in period t, having just observed ω_{jt}, infers that the distribution of ω_{jt+1} is given by $p(\omega_{jt+1}|\omega_{jt})$. Firms thus operate through time, realizing the value of ω_{jt} at period t and forming expectations of future ω_j's using $p(\omega_{jt+1}|\omega_{jt})$. Note that this first-order Markov assumption encompasses the fixed effects assumption where ω_{jt} is fixed over time (i.e. $\omega_{jt} = \omega_j$). OP also assume that $p(\omega_{jt+1}|\omega_{jt})$ is stochastically increasing in ω_{jt}. Intuitively, this means that a firm with a higher ω_{jt} today has a "better" distribution of ω_{jt+1} tomorrow (and in the more distant future). Lastly, note that the ω_{jt} process is assumed to be a time-homogeneous Markov process, i.e. p is not indexed by t.[25]

Third, capital is assumed to be accumulated by firms through a deterministic dynamic investment process, specifically

$$k_{jt} = (1-\delta)k_{jt-1} + i_{jt-1}.$$

Here we will assume that i_{jt-1} is chosen by the firm at period $t-1$. That is, we are assuming that the capital that the firm uses in period t was actually decided upon at period $t-1$; so it takes a full production period for new capital to be ordered, received, and installed by firms.[26] This assumes that capital is a fixed (rather than variable) input.

Lastly, OP specify single period profits as

$$\pi(k_{jt}, a_{jt}, \omega_{jt}, \Delta_t) - c(i_{jt}, \Delta_t).$$

Note that labor l_{jt} is not explicitly in this profit function – the reason is that labor is assumed to be a variable and nondynamic input. It is variable in that (unlike capital), l_{jt} is chosen at period t, the period it actually gets used (and thus it can be a function of ω_{jt}). It is nondynamic in the sense that (again, unlike capital) current choice of labor has no

[25] This assumption is not as strong as it might seem, as, e.g. one can easily allow average productivity to vary across time by indexing β_0 by t, i.e. β_{0t}. The assumption can also be relaxed in some cases, i.e. allowing $p_t(\omega_{jt+1}|\omega_{jt})$ to be indexed by t.

[26] We note that there is a long literature on trying to determine the distributed lag which translates investment expenditures into a productive capital stock [see, for e.g. Pakes and Griliches (1984) and the literature cited there], and one could incorporate different assumptions on this distributed lag into the OP framework. OP themselves also tried allowing current investment to determine current capital, but settled on the specification used here.

impact on the future (i.e. it is not a state variable). This nondynamic assumption rules out, for example, fixed hiring or firing costs of labor. We discuss relaxing this assumption in Section 2.4. For now $\pi(k_{jt}, a_{jt}, \omega_{jt}, \Delta_t)$ can be interpreted as a "conditional" profit function – conditional on the optimal static choice of labor input.

Note also that both $\pi(\cdot)$ and $c(\cdot)$ depend on Δ_t, which represents the economic environment that firms face at a particular point in time. Δ_t could capture input prices, characteristics of the output market, or industry characteristics like the current distribution of the states of firms operating in the industry. The OP formulation allows all these factors to change over time, although they are assumed constant across firms in a given time period. Including market structure in the state space allows some of the competitive richness of the Markov-perfect dynamic oligopoly models of Ericson and Pakes (1995).[27]

Given this economic environment, a firm's maximization problem can be described by the following Bellman equation:

$$\begin{aligned} &V(k_{jt}, a_{jt}, \omega_{jt}, \Delta_t) \\ &\quad = \max\Big\{\Phi(k_{jt}, a_{jt}, \omega_{jt}, \Delta_t), \max_{i_{jt}\geqslant 0}\big\{\pi(k_{jt}, a_{jt}, \omega_{jt}, \Delta_t) - c(i_{jt}, \Delta_t) \\ &\qquad + \beta E\big[V(k_{jt+1}, a_{jt+1}, \omega_{jt+1}, \Delta_{t+1})\big|k_{jt}, a_{jt}, \omega_{jt}, \Delta_t, i_{jt}\big]\big\}\Big\}. \end{aligned}$$

k_{jt}, a_{jt} and ω_{jt} are sufficient to describe the firm specific component of the state space because labor is not a dynamic variable and because $(k_{jt}, a_{jt}, \omega_{jt})$ (and the control i_{it}) are sufficient to describe firms perceived distributions over future $(k_{jt+1}, a_{jt+1}, \omega_{jt+1})$.

The Bellman equation explicitly considers two decisions of firms. First is the exit decision – note that $\Phi(k_{jt}, a_{jt}, \omega_{jt}, \Delta_t)$ represents the sell off value of the firm. Second is the investment decision i_{jt}, which solves the inner maximization problem. Under appropriate assumptions,[28] we can write the optimal exit decision rule as

$$\chi_{jt} = \begin{cases} 1 \text{ (continue)} & \text{if } \omega_{jt} \geqslant \overline{\omega}(k_{jt}, a_{jt}, \Delta_t) = \overline{\omega}_t(k_{jt}, a_{jt}), \\ 0 \text{ (exit)} & \text{otherwise,} \end{cases} \tag{26}$$

and the investment demand function as

$$i_{jt} = i(k_{jt}, a_{jt}, \omega_{jt}, \Delta_t) = i_t(k_{jt}, a_{jt}, \omega_{jt}). \tag{27}$$

[27] See Gowrisankaran (1995), Doraszelski and Satterthwaite (2007), and the third section of this chapter for more discussion of such equilibria.

[28] Other than assuming that an equilibria exists, the main assumption here is that the difference in profits between continuing and exiting is *increasing* in ω_{jt}. Given that ω_{jt} positively affects current profits and that the distribution $p(\omega_{jt+1}|\omega_{jt})$ is stochastically increasing in ω_{jt}, the value of continuing is clearly increasing in ω_{jt}. Thus as long as $\Phi(k_{jt}, \omega_{jt}, \Delta_t)$ either does not depend on ω_{jt}, decreases in ω_{jt}, or does not increase *too fast* in ω_{jt}, this will be satisfied. Note that to get the specific selection bias discussed in Section 2.1 above (i.e. k_{jt} negatively correlated with ω_{jt}), we also need the difference in returns between continuing and exiting to be *increasing* in k_{jt}.

Note the slight change in notation – we are now representing the dependence on Δ_t through the subscript t. See Pakes (1994) for a discussion of conditions under which this investment demand function is strictly increasing in ω_{jt} in the region where $i_{jt} > 0$. That is, conditional on k_{jt} and a_{jt}, firms with higher ω_{jt} optimally invest more. This is an intuitive result – because $p(\omega_{jt+1}|\omega_{jt})$ is assumed stochastically increasing in ω_{jt}, ω_{jt} positively impacts the distribution of all future $\omega_{j\tau}$'s. Since $\omega_{j\tau}$'s positively impact the marginal product of capital in future periods τ, current investment demand should increase. The importance of this *strict monotonicity* condition will be apparent momentarily.

2.3.2. *Controlling for endogeneity of input choice*

Given the setup of the model, we can now proceed with the OP estimation strategy. We first focus on dealing only with the endogeneity of input choice, i.e. we assume there are no selection problems due to exit. We will also assume for now that investment levels are always positive, i.e. $i_{jt} > 0,\ \forall(j,t)$. Later we will relax both these assumptions.

Given that (27) is strictly monotonic in ω_{jt}, it can be inverted to generate

$$\omega_{jt} = h_t(k_{jt}, a_{jt}, i_{jt}). \tag{28}$$

Intuitively, this says that conditional on a firm's levels of k_{jt} and a_{jt}, its choice of investment i_{jt} "tells" us what its ω_{jt} must be. Note that the ability to "invert" out ω_{jt} depends not only on the strict monotonicity in ω_{jt}, but also the fact that ω_{jt} is the *only* unobservable in the investment equation.

This is the *scalar unobservable* assumption mentioned earlier. This, for example, means that there can be no unobserved differences in investment prices across firms,[29] no other state variables that the econometrician does not observe, and no unobserved separate factors that affect investment but not production. It also prohibits ω_{jt} from following higher than a first order Markov process.[30] We discuss both tests for this assumption and the possibilities for relaxing it in Section 2.4.

Substituting (28) into the production function (24) gives

$$y_{jt} = \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + \beta_l l_{jt} + h_t(k_{jt}, a_{jt}, i_{jt}) + \eta_{jt}. \tag{29}$$

The first stage of OP involves estimating (29) using semiparametric methods that treat the inverse investment function $h_t(k_{jt}, a_{jt}, i_{jt})$ nonparametrically. Note the advantages of treating $h_t(k_{jt}, a_{jt}, i_{jt})$ nonparametrically. $i_t(\cdot)$ (and thus its inverse $h_t(\cdot)$) are complicated functions that depend on all the primitives of the model (e.g. demand functions,

[29] Recall that changes in the price of investment over time are permitted as they are picked up by the function h through its dependence on t.

[30] If, for example, ω_{jt} followed a second order process, both ω_{jt} and ω_{jt-1} would enter the state space and the investment decision. With two unobservables in the investment function, it would not be possible to invert out ω_{jt} in the current model.

the specification of sunk costs, the form of conduct in the industry, etc.). These functions are also solutions to a potentially very complicated dynamic game. The OP nonparametric approach therefore avoids both the necessity of specifying these primitives, and the computational burden that would be necessary to formally compute $h_t(\cdot)$.

Given the nonparametric treatment of $h_t(k_{jt}, a_{jt}, i_{jt})$, it is clear that β_0, β_k and β_a cannot be identified using (29). If, for example, $h_t(k_{jt}, a_{jt}, i_{jt})$ is treated as a polynomial in k_{jt}, a_{jt} and i_{jt}, the polynomial will be colinear with the constant, k_{jt}, and a_{jt} terms. Thus, we combine these terms into $\phi_t(k_{jt}, a_{jt}, i_{jt})$, i.e.

$$y_{jt} = \beta_l l_{jt} + \phi_t(k_{jt}, a_{jt}, i_{jt}) + \eta_{jt}. \tag{30}$$

Representing ϕ_t with a high order polynomial in k_{jt}, a_{jt} and i_{jt} [an alternative would be to use kernel methods, e.g. Robinson (1988)] and allowing a different ϕ_t for each time period, OP estimate this equation to recover an estimate of the labor coefficient $\hat{\beta}_l$. To summarize this first stage, the scalar unobservable and monotonicity assumptions essentially allow us to "observe" the unobserved ω_{jt} – this eliminates the input endogeneity problem in estimating the labor coefficient. Note that it is important here that labor is assumed to be a nondynamic input – if labor had dynamic implications, it would enter the state space, and thus the investment function and ϕ_t. As a result, β_l would not be identified in this first stage. Again, this is an assumption that can potentially be relaxed – see Section 2.4.

The second stage of OP identifies the capital and age coefficients β_k and β_a. First, note that the first stage provides an estimate, $\hat{\phi}_{jt}$, of the term

$$\phi_t(k_{jt}, a_{jt}, i_{jt}) = \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + \omega_{jt}.$$

If one uses a polynomial approximation to $\phi_t(k_{jt}, a_{jt}, i_{jt})$, $\hat{\phi}_{jt}$ is just the estimated sum of the polynomial terms for a particular (k_{jt}, a_{jt}, i_{jt}) pair. This means that given a particular set of parameters $(\beta_0, \beta_k, \beta_a)$, we have an estimate of ω_{jt} for all j and t

$$\hat{\omega}_{jt}(\beta_0, \beta_k, \beta_a) = \hat{\phi}_{jt} - \beta_0 - \beta_k k_{jt} - \beta_a a_{jt}. \tag{31}$$

Next decompose ω_{jt} into its conditional expectation given the information known by the firm at $t-1$ (denote this by I_{jt-1}) and a residual, i.e.

$$\begin{aligned}\omega_{jt} &= E[\omega_{jt}|I_{jt-1}] + \xi_{jt} \\ &= E[\omega_{jt}|\omega_{jt-1}] + \xi_{jt} \\ &= g(\omega_{jt-1}) + \xi_{jt}\end{aligned} \tag{32}$$

for some function g. The second line follows from the assumption that ω_{jt} follows an exogenous first order Markov process. By construction, ξ_{jt} is uncorrelated with I_{jt-1}. One can think of ξ_{jt} as the innovation in the ω process between $t-1$ and t that is unexpected to firms. The important thing is that given the information structure of the model, this innovation ξ_{jt} is *by definition* uncorrelated with k_{jt} and a_{jt}. The reason is that k_{jt} and a_{jt} are functions of only the information set at time $t-1$. Intuitively, since

k_{jt} was actually decided on at time $t-1$ (from the investment decision i_{jt-1}), it cannot be correlated with unexpected innovations in the ω process that occurred *after* $t-1$. Lastly, note that since the stochastic process generating ω_{jt} has been assumed constant over time, the g function need not be indexed by t.[31]

Next, consider rewriting the production function as

$$y_{jt} - \beta_l l_{jt} = \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + \omega_{jt} + \eta_{jt}. \tag{33}$$

Substituting in both (32) and (31) results in

$$\begin{aligned} & y_{jt} - \beta_l l_{jt} \\ & \quad = \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + g(\omega_{jt-1}) + \xi_{jt} + \eta_{jt} && \text{(34a)} \\ & \quad = \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + g(\phi_{jt-1} - \beta_0 - \beta_k k_{jt-1} - \beta_a a_{jt-1}) + \xi_{jt} + \eta_{jt} \\ & \quad = \beta_k k_{jt} + \beta_a a_{jt} + \tilde{g}(\phi_{jt-1} - \beta_k k_{jt-1} - \beta_a a_{jt-1}) + \xi_{jt} + \eta_{jt}, && \text{(34b)} \end{aligned}$$

where $\tilde{g}$ encompasses both occurrences of β_0 in the previous line. The key point in (34a) is that, as argued above, the residual $\xi_{jt} + \eta_{jt}$ is uncorrelated with all the right-hand side variables.

We do not observe β_l or ϕ_{jt-1}, but we do have estimates of them from the first stage. Substituting $\hat{\beta}_l$ and $\hat{\phi}_{jt-1}$ for their values in the equation above, and treating $\tilde{g}$ nonparametrically we obtain $\sqrt{n}$ consistent estimates of β_k and β_a. If one uses polynomials to approximate $\tilde{g}$, NLLS can be used for estimation.[32]

Alternatively one can adapt the suggestion in Wooldridge (2004) to combine both stages into a single set of moments and estimate in one step. This should be more efficient than the OP approach (as it uses the information in the covariances of the disturbances, and any cross equation restrictions). The moment condition in this case is

$$E\left[\begin{array}{c} \eta_{jt} \otimes f_1(k_{jt}, a_{jt}, i_{jt}, l_{jt}) \\ (\xi_{jt} + \eta_{jt}) \otimes f_2(k_{jt}, a_{jt}, k_{jt-1}, a_{jt-1}, i_{jt-1}) \end{array}\right] = 0,$$

where f_1 and f_2 are vector valued instrument functions, and $\otimes$ is the Kronecker product operator. Appropriate choices for f_1 and f_2 lead to moments similar to those used by OP. Note that there is a different set of conditioning variables for the moment in η_{jt} than that in the moment for $\xi_{jt} + \eta_{jt}$ (since l_{jt} can be correlated with ξ_{jt}).[33]

[31] Were we to allow $p(\omega_{jt+1}|\omega_{jt})$ to vary across time, we would simply index g by t.

[32] An alternative way to construct a moment condition to estimate (34b) is as follows [see Ackerberg, Caves and Fraser (2004)]. Given β_k and β_a, construct $\hat{\omega}_{jt} = \hat{\phi}_{jt} - \beta_k k_{jt} - \beta_a a_{jt}$, $\forall t$. Non-parametrically regress $\hat{\omega}_{jt}$ on $\hat{\omega}_{jt-1}$ to construct estimated residuals $\hat{\xi}_{jt}$ (note that if using polynomial approximation, this can be done using linear methods (since β_k and β_a are given)). Construct a moment condition interacting $\hat{\xi}_{jt}$ with k_{jt} and a_{jt}. Estimation then involves searching over (β_k, β_a) space to make this moment close to zero.

[33] As Wooldridge notes, one can add further lags of variables to these instrument functions, increasing the number of moments; though more lags will not be able to be used on the observations for the initial years.

2.3.3. Controlling for endogenous selection

Next we relax the assumption that there is no endogenous exit. Firms now exit according to the exit rule given in (26). A first important observation is that the first stage of the OP procedure is not affected by selection. The reason is that by construction, η_{jt}, the residual in the first stage equation (30), represents unobservables that are not observed (or predictable) by the firm before input and exit decisions. Thus there is no selection problem in estimating (30). Intuitively, the fact that in the first stage we are able to completely proxy ω_{jt} means that we can control for both endogenous input choice and endogenous exit.

In contrast, the second stage estimation procedure is affected by endogenous exit. Examining (34b), note that the residual contains not only η_{jt}, but ξ_{jt}. Since the firm's exit decision in period t depends directly on ω_{jt} (see (26)), the exit decision will be correlated with ξ_{jt}, a component of ω_{jt}.[34]

We now correct for the selection. Starting from (33), take the expectation of both sides conditional on both the information at $t-1$ and on $\chi_{jt}=1$ (i.e. being in the dataset at t). This results in

$$\begin{aligned} &E[y_{jt}-\beta_l l_{jt}|I_{jt-1},\chi_{jt}=1] \\ &\quad = E[\beta_0+\beta_k k_{jt}+\beta_a a_{jt}+\omega_{jt}+\eta_{jt}|I_{jt-1},\chi_{jt}=1] \\ &\quad = \beta_0+\beta_k k_{jt}+\beta_a a_{jt}+E[\omega_{jt}|I_{jt-1},\chi_{jt}=1]. \end{aligned} \tag{35}$$

The last line follows because: (1) k_{jt} and a_{jt} are known at $t-1$, and (2) η_{jt} is by definition uncorrelated with either I_{jt-1} or exit at t. Focusing on the last term, we have

$$\begin{aligned} E[\omega_{jt}|I_{jt-1},\chi_{it}=1] &= E\big[\omega_{jt}\big|I_{jt-1},\omega_{jt}\geqslant\overline{\omega}_t(k_{jt},a_{jt})\big] \\ &= \int_{\overline{\omega}_t(k_{jt},a_{jt})}^{\infty}\omega_{jt}\frac{p(\omega_{jt}|\omega_{jt-1})}{\int_{\overline{\omega}_t(k_{jt},a_{jt})}^{\infty}p(\omega_{jt}|\omega_{jt-1})\,\mathrm{d}\omega_{jt}}\,\mathrm{d}\omega_{jt} \\ &= g\big(\omega_{jt-1},\overline{\omega}_t(k_{jt},a_{jt})\big). \end{aligned} \tag{36}$$

The first equality follows from the exit rule. The second and third equalities follows from the exogenous first order Markov process assumption on the ω_{jt} process.

While we do know ω_{jt-1} conditional on the parameters (from (31)), we do not directly observe $\overline{\omega}_t(k_{jt},a_{jt})$. Modelling $\overline{\omega}_t(k_{jt},a_{jt})$ as a nonparametric function of k_{jt}

[34] This correlation relies on OP allowing firms to know the realization of ξ_{jt} before making the exit decision. Otherwise exit would not cause a selection problem. The longer the time period between observations the more serious the selection problem is likely to be. This point comes out clearly in OP's comparison of results based on their "balanced" panel (a data set constructed only from the observations of plants that were active throughout the sample period), to results from their full panel (a panel which keeps the observations on exiting firms until the year they exit and uses observations on new startups from the year they enter). Selection seemed a far larger problem in the balanced than in the full panel.

and a_{jt} might be a possibility, but this would hinder identification of β_k and β_a due to collinearity problems. What we can do is try to control for $\overline{\omega}_t(k_{jt}, a_{jt})$ using data on observed exit. Recall that our exit rule is given by

$$\chi_{jt} = \begin{cases} 1 \text{ (continue)} \\ 0 \text{ (exit)} \end{cases} \quad \text{according as } \omega_{jt} \gtreqless \overline{\omega}_t(k_{jt}, a_{jt}). \tag{37}$$

This means that the probability of being in the data (at period t) conditional on the information known at $t-1$ is

$$\begin{aligned} \Pr(\chi_{jt} = 1 | I_{jt-1}) &= \Pr\big(\omega_{jt} \geqslant \overline{\omega}_t(k_{jt}, a_{jt}) \big| I_{jt-1}\big) \\ &= \Pr\big(\chi_{jt} = 1 \big| \omega_{jt-1}, \overline{\omega}_t(k_{jt}, a_{jt})\big) = \tilde{\tilde{\varphi}}_t\big(\omega_{jt-1}, \overline{\omega}_t(k_{jt}, a_{jt})\big) \\ &= \tilde{\varphi}_t(\omega_{jt-1}, k_{jt}, a_{jt}) = \varphi_t(i_{jt-1}, k_{jt-1}, a_{jt-1}) = P_{jt}. \end{aligned} \tag{38}$$

The second to last equality holds because of (28), and the fact that k_{jt} and a_{jt} are deterministic functions of i_{jt-1}, k_{jt-1}, and a_{jt-1}.

Equation (38) can be estimated nonparametrically, i.e. modelling the probability of surviving to t as a nonparametric function of i_{jt-1}, k_{jt-1}, and a_{jt-1}. OP do this in two alternative ways – first using a probit model with a 4th order polynomial in $(i_{jt-1}, k_{jt-1}, a_{jt-1})$ as the latent index, second using kernel methods. For a plant characterized by $(i_{jt-1}, k_{jt-1}, a_{jt-1})$, these estimates allow us to generate a consistent estimate of the probability of the plant surviving to period t ($\widehat{P}_{jt}$).

Next, note that as long as the density of ω_{jt} given ω_{jt-1} is positive in an area around $\overline{\omega}_t(k_{jt}, a_{jt})$, (38) can be inverted to write $\overline{\omega}_t(k_{jt}, a_{jt})$ as a function of ω_{jt-1} and P_{jt},[35] i.e.

$$\overline{\omega}_t(k_{jt}, a_{jt}) = f(\omega_{jt-1}, P_{jt}). \tag{39}$$

Substituting (39) into (36) and (35), and using (31) gives us

$$\begin{aligned} E[y_{jt} - \beta_l l_{jt} | I_{jt-1}, \chi_{jt} = 1] &= \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + g\big(\omega_{jt-1}, f(\omega_{jt-1}, P_{jt})\big) \\ &= \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + g'(\omega_{jt-1}, P_{jt}) \\ &= \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + g'(\phi_{jt-1} - \beta_0 \\ &\quad - \beta_k k_{jt-1} - \beta_a a_{jt-1}, P_{jt}). \end{aligned} \tag{40}$$

This is similar to (35), only differing in the additional P_{jt} term in the nonparametric g' function. P_{jt} controls for the impact of selection on the expectation of ω_{jt} – i.e. firms with lower survival probabilities who *do in fact* survive to t likely have higher ω_{jt}'s than those with higher survival probabilities.

[35] Formally, (38) implies that $P_{jt} = \tilde{\tilde{\varphi}}_t(\omega_{jt-1}, \overline{\omega}_{jt})$. With positive density of ω_{jt} around $\overline{\omega}_{jt}$, $\tilde{\tilde{\varphi}}_t$ is strictly monotonic in $\overline{\omega}_{jt}$, so this can be inverted to generate (39).

Equation (40) implies that we can write

$$\begin{aligned} y_{jt} - \beta_l l_{jt} &= \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + g'(\phi_{jt-1} - \beta_0 - \beta_k k_{jt-1} - \beta_a a_{jt-1}, P_{jt}) + \zeta_{jt} \\ &= \beta_k k_{jt} + \beta_a a_{jt} + \tilde{g}(\phi_{jt-1} - \beta_k k_{jt-1} - \beta_a a_{jt-1}, P_{jt}) + \zeta_{jt} + \eta_{jt}, \end{aligned} \tag{41}$$

where, as in (34b), the two β_0 terms have been encompassed into the nonparametric function $\tilde{g}$. By construction the residual in this equation satisfies $E[\zeta_{jt} + \eta_{jt} | I_{jt-1}, \chi_{jt} = 1] = 0$. Substituting $\widehat{P}_{jt}$, $\hat{\phi}_{jt}$ and $\hat{\beta}_l$ for P_{jt}, ϕ_{jt} and β_l, (41) can also be estimated with NLLS, approximating $\tilde{g}$ with either a polynomial or a kernel.[36]

In this estimation procedure information on β_k and β_a is obtained by comparing labor productivities of firms with the same ω_{jt-1} and P_{jt} but different k_{jt} and a_{jt}. In addition, since the functions $\varphi_t(\cdot)$ and $P_t(\cdot)$ vary across t with changes in industry conditions (while $g(\cdot)$ is assumed constant over time), it also uses information from variation in inputs across firms in *different* time periods that have the same ω_{jt-1} and P_{jt}.

In the selection literature, $\widehat{P}_{jt}$ is referred to as the propensity score – for discussion of these techniques, see, e.g. Heckman (1974, 1978, 1979), Rosenbaum and Rubin (1983), Heckman and Robb (1985), and Ahn and Powell (1993). An important difference between OP and this selection literature is that controlling for the propensity score is not sufficient for OP's model; they require a control for both ω_{jt-1} and for $\overline{\omega}_{jt-1}$.

A number of recent papers have applied the OP procedure successfully. As an example consider Table 3, which displays the results from the food processing industry in Pavcnik (2002) – this is the first out of the eight industries in her Table 2. Comparing the OLS to the OP estimates, we see the changes that we should expect. Returns to scale decrease (consistent with positive correlation between unobserved productivity and input use), with the coefficients on the more variable inputs accounting for all of the fall (consistent with this correlation being more pronounced for the variable inputs). Consistent with selection, the capital coefficient rises moving from OLS to OP. The fixed effects estimates are the most difficult to understand, as they generate a coefficient for capital near zero, and an estimate of economies of scale below 0.9. These results are indicative of those for the other industries in Pavcnik's Table 2. The average of the returns to scale estimate across industries when estimated by OLS is 1.13, when estimated by OP it is 1.09, and when estimated by fixed effects it is 0.87. The average of the capital coefficients across industries from OLS is 0.066, from OP 0.085, and from fixed effects only 0.021 (with two industries generating negative capital coefficients).

OP themselves compare their estimates to estimates obtained using OLS and fixed effect on both a balanced panel (a panel constructed only from firms that were operating during the entire fifteen year sample period) and from the full sample (constructed by keeping firms that eventually exit until the year prior to their exit and introducing new

[36] OP try both the kernel and a polynomial with only minor differences in results.

Table 3
Production function estimates from Pavcnik (2002)

	OLS	Fixed effects	Olley–Pakes
Unskilled labor	0.178	0.210	0.153
	(0.006)	(0.010)	(0.007)
Skilled labor	0.131	0.029	0.098
	(0.006)	(0.007)	(0.009)
Materials	0.763	0.646	0.735
	(0.004)	(0.007)	(0.008)
Capital	0.052	0.014	0.079
	(0.003)	(0.006)	(0.034)

Source: From Pavcnik (2002).

entrants as they appear). The difference between the balanced sample estimators and OP estimators on the full sample are truly dramatic, and those between the OLS and fixed effect estimators on the full sample and the OP estimators are similar to those reported above (though somewhat larger in absolute value). In both papers, the OP estimator generates standard errors for the labor coefficient that are not too different then those generated by OLS, but, as might be expected, standard errors for the capital coefficient do increase (though much less so in the OP results than in Pavcnik's).

2.3.4. Zero investment levels

For simplicity, we assumed above that investment levels for all observations were nonzero. This allowed us to assume that the investment equation was strictly monotonic in ω_{jt} everywhere (and hence could be inverted to recover ω_{it} for every observation). Observations with zero investment call into question the strict monotonicity assumption. However, the OP procedure actually only requires investment to be strictly monotonic in ω_{jt} for a *known* subset of the data. OP themselves take that subset to be all observations with $i_t > 0$, i.e. they simply do not use the observations where investment equals 0.

Even with this selected sample, first stage estimation of (29) is consistent. Since ω_{jt} is being completely proxied for, the only unobservable is η_{jt}, which is by assumption uncorrelated with labor input and with the selection condition $i_{it} > 0$. Second stage estimation of (41) is also consistent when OP discard the data where $i_{jt-1} = 0$ ($\hat{\phi}_{jt-1} - \beta_0 - \beta_k k_{jt-1} - \beta_a a_{jt-1}$ is not computable when $i_{jt-1} = 0$). The reason is that the error term in (41) is by construction uncorrelated with the information set I_{jt-1}, which contains the investment level i_{jt-1}. In other words, conditioning on $i_{jt-1} = 0$ does not say anything about the unobservable ζ_{jt}.

While the OP procedure can accommodate zero investment levels, this accommodation is not without costs. In particular, there is likely to be an efficiency loss from discarding the subset of data where $i_{jt} > 0$. Levinsohn and Petrin (2003) (henceforth

LP) suggest an alternative estimation routine whose primary motivation is to eliminate this efficiency loss. They start by noting that in many datasets, particularly those from developing countries, the set of observations with zero investment can be quite large. For example, in LP's dataset on Chilean plants more than 50% of the observations have zero investment (note that in OP's US plant data, this proportion is much less, ≈8%). To avoid a potentially large efficiency loss, LP suggest using variables other than investment to proxy for the unobserved ω_{jt}. In particular, LP focus on firms' choices of intermediate inputs (e.g. electricity, fuels, and/or materials) – these are rarely zero.[37]

Consider the production function

$$y_{jt} = \beta_0 + \beta_k k_{jt} + \beta_l l_{jt} + \beta_m m_{jt} + \omega_{jt} + \eta_{jt} \tag{42}$$

with additional input m_{jt} (e.g. materials). LP assume that like labor, m_{jt} is a variable (i.e. chosen at t), nondynamic input, and consider the following material demand equation

$$m_{jt} = m_t(k_{jt}, \omega_{jt}). \tag{43}$$

As with the OP investment equation, the demand equation is indexed by t to allow, e.g. input prices, market structure, and demand conditions to vary across time.[38] LP state conditions under which this demand equation is monotonic in ω_{jt}. Given this monotonicity, estimation proceeds analogously to OP. First, (43) is inverted to give

$$\omega_{jt} = h_t(k_{jt}, m_{jt}). \tag{44}$$

Next, (44) is substituted into (42) to give

$$y_{jt} = \beta_0 + \beta_k k_{jt} + \beta_l l_{jt} + \beta_m m_{jt} + h_t(k_{jt}, m_{jt}) + \eta_{jt}. \tag{45}$$

Treating the h_t function nonparametrically results in the following estimating equation

$$y_{jt} = \beta_l l_{jt} + \phi_t(k_{jt}, m_{jt}) + \eta_{jt}, \tag{46}$$

where β_k and β_m are not separately identified from the nonparametric term. As in OP, the first stage of LP involves estimating (46) to obtain $\hat{\beta}_l$ and $\hat{\phi}_{jt}$. The second stage of LP again proceeds following OP, the main difference being that the parameter on the intermediate input, β_m, still needs to be estimated. Moving the labor term to the

[37] An alternative to LP might be to augment the original OP procedure with a more complete model of investment and/or distributional assumptions on ω, allowing one to utilize the zero investment observations.

[38] Given that materials are a static choice (in contrast to dynamic investment), one might be more willing to make parametric assumptions on this input demand function (since it depends on fewer primitives, e.g. it does not depend on expectations about the future). However, there are caveats of such an approach, see Section 2.4.1.

left-hand side and using (32) gives[39]

$$\tilde{y}_{jt} = \beta_k k_{jt} + \beta_m m_{jt} + \tilde{g}(\phi_{jt-1} - \beta_k k_{jt-1} - \beta_m m_{jt-1}) + \xi_{jt} + \zeta_{jt}, \tag{47}$$

and nonparametric estimates of ϕ_{jt} and of $\tilde{g}(\cdot)$ are used in estimation.

Note that since k_{jt} is assumed decided at $t-1$, it is orthogonal to the residual, $\xi_{jt} + \eta_{jt}$. However, since m_{jt} is a variable input, it is clearly not orthogonal to ζ_{jt}, the innovation component of ω_{jt}. LP address this by using m_{jt-1} as an instrument for m_{jt} in estimation of (47). In their application LP find biases that are generally consistent with those predicted by OP, but some differences in actual magnitudes of coefficients.

2.4. Extensions and discussion of OP

The OP model was designed to produce estimates of production function coefficients which are not subject to biases due to simultaneity and selection problems generated by the endogeneity of input demands and exit decisions. We begin this section with a test of whether the coefficient estimates obtained using OP's assumptions are robust to different sources of misspecification.

There are a variety of reasons why this test could fail and the rest of this subsection considers some of the more likely candidates. Each time a source of possible misspecification in OP's assumption is introduced, we consider modifications to their estimation techniques which produce consistent estimates of production function coefficients under that misspecification. This is in keeping with our belief that different modifications are likely to be appropriate for different industries and data sets. Though the extended models may well be of more general interest, as they typically will produce richer dynamics with more detailed policy implications, we limit ourselves to considering their implications for estimating production function coefficients.

In this context we first investigate relaxing assumptions on the dynamic implications of inputs (e.g. that labor choices today have no dynamic implications) and on the timing of input choices. We then investigate the potential for relaxing the scalar unobservable assumptions of OP. Most of the discussion regarding the timing and dynamic implications of inputs is based on Ackerberg, Caves and Fraser (2004) (ACF) [also see Buettner (2004a) for some related ideas], while much of the discussion on nonscalar unobservables is taken from Ackerberg and Pakes (2005). We also briefly discuss two recent contributions by Buettner (2004b) and Greenstreet (2005).

2.4.1. A test of Olley and Pakes' assumptions

This subsection combines results from Section 4.1 in OP with results from ACF. Broadly speaking, there are two questionable implications of the assumptions used in OP that

[39] While the LP procedure does not formally address selection, they note that their procedure could be extended to control for it in the same way as OP.

are central to their estimation strategy. First there is the implication that, conditional on capital and age, there is a one to one mapping between investment and productivity (we give reasons for doubting this implication below). Second there is the direct assumption that the choice of labor has no dynamic implications; i.e. that labor is not a state variable in the dynamic problem.

Focusing on the second assumption first, assume instead that there are significant hiring or firing costs for labor, or that labor contracts are long term (as in, for example, unionized industries). In these cases, current labor input choices have dynamic implications, labor becomes a state variable in the dynamic problem, and Equation (28) becomes

$$\omega_{jt} = h_t(k_{jt}, l_{jt}, a_{jt}, i_{jt}). \tag{48}$$

Now the labor coefficient will not be identified in the first stage; i.e. from Equation (34b) – the first stage cannot separate out the impact of labor on production, or β_l, from its impact on the $h(\cdot)$ function.

ACF point out that under these assumptions β_l can still be identified from the second stage. To see this note that the second stage is now

$$\begin{aligned} y_{jt} = {} & \beta_l l_{jt} + \beta_k k_{jt} + \beta_a a_{jt} + \tilde{g}(\phi_{jt-1} - \beta_l l_{jt-1} - \beta_k k_{jt-1} - \beta_a a_{jt-1}) \\ & + \xi_{jt} + \eta_{jt}. \end{aligned} \tag{49}$$

After substituting $\hat{\phi}_{jt-1}$ for ϕ_{jt-1}, we can estimate the production function parameters using a semiparametric GMM procedure related to the above. Note, however, that if we maintain the rest of OP's assumptions, then l_{jt} differs from k_{jt} in that labor can adjust to within period variation in productivity. This implies that unlike k_{jt}, l_{jt} can be correlated with ξ_{jt}. As a result we need to use an "instrument" for l_{jt} when estimating Equation (49). A fairly obvious instrument is l_{jt-1}. Since l_{jt-1} was decided on at $t-1$, it is uncorrelated with ξ_{jt}, and l_{jt} and l_{jt-1} are typically highly correlated. With this modification, estimation can proceed as before using, say, a polynomial or kernel approximation to $\tilde{g}$.

Note that even though the first stage does not directly identify any of the parameters of the model in this procedure, we still need the first stage to generate estimates of $\hat{\phi}_{jt-1}$. Indeed we still need (an extended version) of the assumptions that generates the first stage equation. Before we needed the assumption that conditional on values for (k_{jt}, a_{jt}) there was a one to one map between productivity and investment. Now we need the assumption that conditional on values of (k_{jt}, a_{jt}, l_{jt}) there is a one to one map between productivity and investment.

In fact Equation (49) is closely related to the test for the inversion proposed in OP. Recall that they assume that labor is not a dynamic input. In that case when they subtract their first stage estimate $\hat{\beta}_l$ times l from both sides of their second stage equation they obtain

$$y_{jt} - \hat{\beta}_l l_{jt} = (\beta_l - \hat{\beta}_l) l_{jt} + \beta_k k_{jt} + \beta_a a_{jt} + \tilde{g}(\hat{\phi}_{jt-1} - \beta_k k_{jt-1} - \beta_a a_{jt-1}) + \xi_{jt} + \eta_{jt}, \tag{50}$$

which is an equation with over identifying restrictions.[40]

In particular, the term $(\beta_l - \hat{\beta}_l) l_{jt}$ in Equation (50) should be zero if the inversion which leads to the estimate of the labor coefficient is a good approximation to reality. Further the inversion implies that what we must subtract from our estimate of ϕ_{jt-1} to obtain lagged productivity is determined by the contribution of (k_{jt-1}, a_{jt-1}) to production of y_{jt-1}, i.e. by (β_k, β_a). These coefficients also determine the contribution of (k_{jt}, a_{jt}) to y_{jt} given ω_{jt-1}. If our inversion were seriously in error we would expect that $\phi_{jt-1} - \beta_k k_{jt-1} - \beta_a a_{jt-1}$ would not be perfectly correlated with ω_{jt-1} and as a result there would be a residual component of productivity we are not controlling for. Provided there was an endogeneity problem in the first place, this residual should be correlated with (k_{jt}, a_{jt}). Thus OP allow the coefficients on (k_{jt}, a_{jt}) to differ from those on (k_{jt-1}, a_{jt-1}) in Equation (50), use $(l_{j,t-1}, k_{jt}, a_{jt})$ and powers and lags of these variables as instruments, and then test whether $\hat{\beta}_l - \beta_l = 0$, and whether the coefficients on the current and lagged values of k and a are equal.[41]

As noted previously, when the current labor choice has dynamic implications the first stage estimate of β_l obtained by OP is inconsistent (regardless of whether the inversion is correct). However even if labor is dynamic, Equation (49) still generates over identifying restrictions; the existence of the inversion implies that the current and lagged values of (l, k, a) should enter in this equation with the same factors of proportionality. In other words, if the inversion is correct then what we must subtract from our estimate of ϕ_{jt-1} to obtain lagged productivity is determined by the contribution of $(l_{jt-1}, k_{jt-1}, a_{jt-1})$ to production of y_{jt-1}, i.e. by $(\beta_l, \beta_k, \beta_a)$. These coefficients also determine the contribution of (l_{jt}, k_{jt}, a_{jt}) to y_{jt} given ω_{jt-1}. That is if we were to estimate

$$y_{jt} = \beta_l^* l_{jt} + \beta_k^* k_{jt} + \beta_a^* a_{jt} + \tilde{g}(\phi_{jt-1} - \beta_l l_{jt-1} - \beta_k k_{jt-1} - \beta_a a_{jt-1}) + \xi_{jt} + \eta_{jt}, \tag{51}$$

and then test whether $(\beta_l^*, \beta_k^*, \beta_a^*) = (\beta_k, \beta_l, \beta_a)$, with a large enough data set we should reject the null of equality if assumptions which underlie the inversion are wrong. Given the additional parameters one will need additional instruments to estimate this specification. Natural instruments would be those used in OP, i.e. $(l_{j,t-1}, k_{jt}, a_{jt})$ and powers and lags of these variables.

Two other points about the test. First, the OP test conditions on the fact that labor is variable (i.e. it is not a state variable) and endogenous (current labor is correlated

[40] We have omitted a term that results from substituting $\hat{\phi}_{jt-1}$ for the true ϕ_{jt-1} in this equation. The additional term's impact on the parameter estimates is $o_p(1/\sqrt{J})$, and so does not effect their limit distributions.

[41] Note that we cannot use both current and lagged values of a_{jt} as instruments for the two are collinear. We could, however, use different functions of a_{jt} as additional instruments.

with ξ_{jt}), and then tests whether the inversion is a good approximation. We could have alternatively proceeded by conditioning on the inversion and then tested one or both of the assumptions that; labor is dynamic and/or labor choices are fixed prior to the realization of ξ_t. We would do this by estimating both stages of ACF simultaneously and then testing constraints. The constraint to be tested in asking whether labor can be treated as a nondynamic (or variable) input is whether $\phi(l_{jt}, i_{jt}, k_{jt}, a_{jt}) = \beta_l l_{jt} + \phi(i_{jt}, k_{jt}, a_{jt})$. To test whether labor is endogenous (in the sense that it can react to ξ_{jt}) we estimate the system once using l_{jt-1} as an instrument for l_{jt} in Equation (49) and once using l_{jt} as an instrument for itself. Exactly what it makes sense to condition on (and what to test for) is likely to depend on the characteristics of the industry being studied. Alternatively we could improve the power of the omnibus test in Equation (49) by estimating the first stage in ACF simultaneously with this equation and then asking whether $(\beta_l^*, \beta_k^*, \beta_a^*) = (\beta_l, \beta_k, \beta_a)$. If that is accepted we could then test the additional (nested) constraints implied by an assumption that labor is not endogeous.[42]

Finally a word of caution on the usefulness of these tests. First we have made no attempt to look at the power of these tests. Though OP find very precise estimates of differences in coefficients from (51), their data seems to deliver more precise estimates than many other data sets (see, for e.g. ACF). Second it is important to realize that the test that $(\beta_l^*, \beta_k^*, \beta_a^*) = (\beta_l, \beta_k, \beta_a)$ is designed to ask the limited question of whether making our approximations greatly hinders our ability to obtain reasonable production function coefficients. As a result we are using the difference in these coefficients, normalized by the variance-covariance of those differences, as our metric for "reasonableness". There are other metrics possible, one of which would be to have some prior knowledge of the characteristics of the industry the researcher is working on (and we expect these results to vary by industry). Further there may well be independent reasons for interest in the timing of input decisions or in our invertibility assumption (see the discussion below), and a test result that our approximations do not do terrible harm to production function estimates does not imply that they would do little harm in the analysis of other issues (for example in the analysis of the response of labor hiring to a change in demand, or in the response of investment to an infrastructure change which increases productivity).

[42] Note that both these ideas: that one can allow labor to have dynamic effects and that some of the assumptions behind these procedures are testable – are related to the dynamic panel literature cited above [e.g. Arellano and Bond (1991), Arellano and Bover (1995), and Blundell and Bond (1999)] in that further lags of inputs are typically used as instruments. If one were willing to assume that the $\eta_{j,t}$ are independently distributed across time then the residuals should be uncorrelated with past values of output also. However if η_{jt} represented serially correlated measurement error in the observations on y_t then the η_{jt} may be serially correlated, and we could not expect a zero correlation between past output and the disturbance from (51). ACF flesh out the distinction between their methods and the dynamic panel literature further.

2.4.2. Relaxing assumptions on inputs

This subsection assumes that there is an inversion from productivity to investment conditional on the state variables of the problem, and investigates questions regarding the nature of the input demands given this inversion. ACF note that there are two dimensions along which we can classify inputs in this context, and the two dimensions have different implications for the properties of alternative estimators. First inputs can either be variable (correlated with ξ_{jt}) or fixed (uncorrelated with ξ_{jt}). Second the inputs can either be dynamic, i.e. be state variables in the dynamic problem and hence conditioned on in the relationship between productivity and investment, or static. So if we generalize, and allow for inputs of each of the four implicit types we have

$$y_{jt} = \beta^{\text{vs}} X_{jt}^{\text{vs}} + \beta^{\text{vd}} X_{jt}^{\text{vd}} + \beta^{\text{fs}} X_{jt}^{\text{fs}} + \beta^{\text{fd}} X_{jt}^{\text{fd}} + \omega_{jt} + \eta_{jt}, \tag{52}$$

where the input acronyms correspond to these dimensions, e.g. X_{jt}^{vs} represent variable, nondynamic inputs, while X_{jt}^{fd} represent fixed, dynamic inputs, and so on.

The various coefficients can be identified in different ways. β^{vs}, like labor in the original OP framework, can be identified either in the first stage, or in the second stage using X_{jt-1}^{vs} as an instrument (because X_{jt}^{vs} is variable and thus potentially correlated with ξ_{jt}, it cannot be used as an instrument in the second stage). β^{fd}, like capital in the original OP framework, cannot be identified in the first stage, but it can be identified in the second stage using either X_{jt}^{fd} or X_{jt-1}^{fd} (or both) as instruments. β^{vd}, the coefficients on the inputs that are variable and dynamic, also cannot be identified in the first stage, but can be identified in the second stage using X_{jt-1}^{vd} as an instrument. Lastly, β^{fs} can be identified either in the first stage or in the second stage using either X_{jt}^{fs} or X_{jt-1}^{fs} (or both) as instruments.

Note also that if we have any static or fixed inputs we have over identifying restrictions.[43] This over identification can potentially be useful in testing some of the timing assumptions. For example, suppose one starts by treating capital as a fixed, dynamic input. One could then estimate the second stage using both k_{it-1} and k_{it} as instruments, an over identified model. In the GMM context, one could test this over identification with a J-test [Hansen (1982)]. Since k_{it} is a valid instrument only when capital is truly fixed (yet k_{it-1} is a valid instrument regardless) rejection of the specification might be interpreted as evidence that capital is not a completely fixed input. Consistent estimation could then proceed using only k_{it-1} as an instrument. Again, the Wooldridge (2004) framework makes combining these multiple sources of identification and/or testing very convenient.

ACF also look deeper into the various assumptions on inputs. They note that, under the assumption that l_{jt} is a variable input, for it to have the independent variance needed

[43] In all these cases, further lags (prior to $t-1$) of the inputs can be used as instruments and thus as over identifying restrictions, although it is not clear how much extra information is in these additional moment conditions, and one will not be able to use these additional lags in the initial time periods.

to estimate our first stage equation (30), there must be a variable, say z_{jt}, that impacts firms' choices of l_{jt} but that does not impact choices of investment at t. This variable z_{jt} must also have some variance that is independent of ω_{jt} and k_{jt}. If this were not the case, e.g. if $l_{jt} = f_t(\omega_{jt}, k_{jt})$, then one can show that l_{jt} is perfectly collinear with the nonparametric function in Equation (30), implying that one cannot estimate β_l from that equation. Note that the variable z_{jt} does not need to be observed by the econometrician.

Thus, to proceed as OP do we need the demand function for labor to be

$$l_{jt} = f_t(\omega_{jt}, k_{jt}, z_{jt}),$$

where z_{jt} are additional factors that affect demand for labor (or more generally, demand for the variable inputs) with nonzero conditional variance (conditional on ω_{jt}, k_{jt}). Note that the z_{jt} *cannot* be serially correlated. If this were the case, then z_{jt} would become part of the state space, influence i_{jt}, and one would not be able to do the inversion.[44]

Even with this restriction, there are at least two possible z_{jt}'s here: i.i.d. firm specific input price shocks and i.i.d. random draws to the environment that cause differences in the variance of η_{jt} over time (since the profit function is a convex function of η the variance in this variable will affect labor demand). The latter could be associated with upcoming union negotiations, the likelihood of machine break downs due to age of equipment, or the approach of maintenance periods. One problem with the i.i.d. input price shock story is that it is somewhat at odds with the assumptions that all other components of prices are constant across firms and that the other unobservables (ω_{jt}) in the model are serially correlated over time.

ACF provide two additional ways of overcoming this problem. First they note that if one weakens OP's timing assumptions slightly, one can still identify l_{jt} in the first stage. Their observation also reopens an avenue of research on the timing of input decisions which dates back at least to Nadiri and Rosen (1974). Suppose that l_{jt} is actually not a perfectly variable input, and is chosen at some point in time between period $t-1$ and period t. Denote this point in time as $t-b$, where $0 < b < 1$. Suppose that ω evolves between the subperiods $t-1$, $t-b$, and t according to a first order Markov process, i.e.

$$p(\omega_{jt}|I_{jt-b}) = p(\omega_{jt}|\omega_{jt-b}) \quad \text{and} \quad p(\omega_{jt-b}|I_{jt-1}) = p(\omega_{jt-b}|\omega_{jt-1}).$$

In this case, labor input is not a function of ω_{jt}, but of ω_{jt-b}, i.e.

$$l_{jt} = f_t(\omega_{jt-b}, k_{jt}).$$

Since ω_{jt-b} cannot generally be written as a function of k_{jt}, a_{jt}, and i_{jt}, l_{jt} will *not* generally be collinear with the nonparametric term in (30), allowing the equation to be identified. The movement of ω between $t-b$ and t is what breaks the collinearity problem between l_{jt} and the nonparametric function. The second alternative suggested

[44] Note also that observing z_{jt} would not help in this serially correlated case. While one would now be able to do the inversion, z_{jt} would enter the nonparametric function, again generating perfect collinearity.

by ACF avoids this collinearity problem by abandoning the first stage identification of the labor coefficient. Instead, they suggest identifying the labor coefficient in the second stage using l_{jt-1} as an instrument.

Importantly, ACF argue that this collinearity problem is more severe when using the LP procedure. They contend that it is considerably harder to tell a believable story in which the assumptions of LP hold and where l_{jt} varies independently of the nonparametric function in (46). The reason for this is that it is hard to think of a variable z_{jt} that would affect a firms' labor choices but not their material input choices (either directly or indirectly through the labor choice).[45] ACF suggest a couple of procedures as alternatives to LP.[46] The first, based on the discussion above, again involves simply identifying the labor coefficient in the second stage. This can be done using either l_{jt} or l_{jt-1} to form an orthogonality condition, depending on what one wants to assume about the timing of the labor choice. Moreover, it can also be done in a manner that is also consistent with labor having dynamic effects. The second procedure is more complicated and involves sequentially inverting the value of ω_{jt} at each point in time at which inputs are chosen. While this procedure depends on independence (rather than mean independence) assumptions on innovations in ω_{jt}, it has the added advantage of allowing one to infer something about the point in time that labor is chosen. Bond and Söderbom (2005) make a somewhat related point regarding collinearity. They argue that in a Cobb–Douglas context where input prices are constant across firms, it is hard if not impossible to identify coefficients on inputs that are perfectly variable and have no dynamic effects. This is important for thinking about identification of coefficients on X_{jt}^{vs} in the above formulation.

2.4.3. Relaxing the scalar unobservable assumption

The assumption of a scalar unobserved state variable is another aspect of the OP approach that might be a source of concern. We begin with three reasons for worrying about this assumption and then provide a way of modifying the model to account for each of them. In each case we bring information on additional observables to bear on the problem. As a result, one way of looking at this section is as a set of robustness tests conducted by asking whether the additional observables affect the results.

Our three concerns in order of increasing difficulty are as follows. First productivity itself is a complex functions of many factors, and it may not be appropriate to assume

[45] ACF note that one probably will not observe this perfect collinearity problem in practice (in the sense that the first stage procedure will actually produce an "estimate"). However, they point out that unless one is willing to make what they argue are extremely strong and unintuitive assumptions, the lack of perfect collinearity in practice must come from misspecification in the LP model.

[46] An alternative approach to dealing with these collinearity problems might be to model the input demand functions (investment or materials) parametrically. If $g()$ is parametric, one does not necessarily have this collinearity problem. However, at least in the LP situation this does not guarantee identification. ACF show that in the Cobb–Douglas case, substituting in the implied parametric version of the material input function leads to an equation that cannot identify the labor coefficient.

that one can represent it as a first order Markov process. Second investment might well respond to demand factors that are independent of the firm's productivity. Then there is no longer a one to one mapping between investment and productivity given capital and age. Consequently we cannot do the inversion in Equation (28) underlying the first stage of the OP procedure. Finally, at least in some industries we often think of two sources of increments in productivity, one that results from the firm's own research investments, and one whose increments do not depend on the firm's behavior. A process formed from the sum of two different first order Markov processes is not generally a first order Markov process, and if one of those processes is "controlled" it may well be difficult to account for it in the same way as we can control for exogenous Markov processes.

First assume that productivity follows a second order (rather than first order) Markov process. This changes the investment demand equation to

$$i_{jt} = i_t(k_{jt}, a_{jt}, \omega_{jt}, \omega_{jt-1}). \tag{53}$$

Since there are the two unobservables $(\omega_{jt}, \omega_{jt-1})$ the investment equation cannot be inverted to obtain ω_{jt} as a function of observables, and the argument underlying the first stage of the OP process is no longer valid.

One possible solution to the estimation problem is through a second observed control of the firm. Suppose, for example, one observes firms' expenditures on another investment (advertising, expenditure on a distributor or repair network), say s_{it}.[47] Then we have the bivariate policy function

$$\begin{pmatrix} i_{jt} \\ s_{jt} \end{pmatrix} = \Upsilon_t(k_{jt}, a_{jt}, \omega_{jt}, \omega_{jt-1}).$$

If the bivariate function $\Upsilon_t \equiv (\Upsilon_{1,t}, \Upsilon_{2,t})$ is a bijection in $(\omega_{jt}, \omega_{jt-1})$ (i.e. it is onto), then it can be inverted in ω_{jt} to obtain

$$\omega_{jt} = \Upsilon_t^{-1}(k_{jt}, a_{jt}, i_{jt}, s_{jt}).$$

Given this assumption the first stage proceeds as in OP, except with a higher dimensional nonparametric function to account for current productivity (it is a function of s_{jt} as well as (k_{jt}, a_{jt}, i_{jt})).

OP's second stage is modified to be

$$\begin{aligned} \tilde{y}_{jt} = {} & \beta_k k_{jt} + \beta_a a_{jt} \\ & + \tilde{g}(\hat{\phi}_{jt-1} - \beta_k k_{jt-1} - \beta_a a_{jt-1}, \hat{\phi}_{jt-2} - \beta_k k_{jt-2} - \beta_a a_{jt-2}) \\ & + \xi_{jt} + \eta_{jt}, \end{aligned}$$

where $\tilde{y}_{jt} = y_{jt} - \hat{\beta} l_{jt}$ and the $\hat{\phi}_{jt}$ variables are obtained from the first stage estimates at $t-1$ and $t-2$. Note that since the conditional expectation of ω_{jt} given I_{jt-1} now

[47] One can modify this argument to allow also for a second state variable, the stock of advertising or the size of the repair network, provided that stock is known up to a parameter to be estimated.

depends on ω_{jt-2} as well as ω_{jt-1}, we need to use estimates of ϕ from two prior periods. The extension to control for selection as well is straightforward. Moreover, provided the number of observed control variables is at least equal to the order of the Markov process, higher order Markov processes can be handled in the same way.

We now move on to allow investment to depend on an unobservable demand shock that varies across firms, in addition to the (now first order) ω_{jt} process. Suppose that the demand shock, μ_{jt}, also follows a first order Markov process that is *independent* of the ω_{jt} process. Then the investment function will be a function of both unobservables, or $i_{jt} = i_t(k_{jt}, a_{jt}, \omega_{jt}, \mu_{jt})$. Again we will assume the existence of a second control and use it to allow us to substitute for ω_{jt} in the first stage of OP's procedure.

More precisely, assume we also observe the firms' pricing decisions p_{jt}. At the risk of some notational confusion, again let the bivariate policy function determining (i_{jt}, p_{jt}) be labelled $\Upsilon(\cdot)$, and assume it is a bijection in (ω_{jt}, μ_{jt}) conditional on (k_{jt}, a_{jt}). Then it can be inverted to form

$$\omega_{jt} = \Upsilon_t^{-1}(k_{jt}, a_{jt}, i_{jt}, p_{jt}) \tag{54}$$

and one can proceed with the first stage of estimation as above.

For the second stage observe that since the μ_{jt} process is independent of the ω_{jt} process the firm's conditional expectation of ω_{jt} given I_{jt-1} only depends on ω_{jt-1}. Thus, the second stage is

$$\tilde{y}_{jt} = \beta_k k_{jt} + \beta_a a_{jt} + \tilde{g}(\hat{\phi}_{jt-1} - \beta_k k_{jt-1} - \beta_a a_{jt-1}) + \xi_{jt} + \eta_{jt}. \tag{55}$$

Note that the demand shock, if an important determinant of i_{jt}, may help with the precision of our estimates, as it generates independent variance in $\hat{\phi}$.

The estimation problem becomes more complicated if, for some reason, the two Markov processes are dependent. The problem is that in this case, the firm's conditional expectation of ω_{jt} given I_{jt-1} depends on *both* ω_{jt-1} and μ_{jt-1}. Then Equation (55) will have to be amended to allow $\tilde{g}(\cdot)$ to also depend on μ_{jt-1}. If we let

$$\mu_{jt-1} = \Upsilon_{2,t-1}^{-1}(k_{jt-1}, a_{jt-1}, i_{jt-1}, p_{jt-1}), \tag{56}$$

our second stage can then be written as

$$\begin{aligned}\tilde{y}_{jt} &= \beta_k k_{jt} + \beta_a a_{jt} + \tilde{g}(\omega_{jt-1}, \mu_{jt-1}) + \xi_{jt} + \eta_{jt} \\ &= \beta_k k_{jt} + \beta_a a_{jt} \\ &\quad + \tilde{g}\big(\phi_{jt-1} - \beta_k k_{jt-1} - \beta_a a_{jt-1}, \Upsilon_{2,t-1}^{-1}(k_{jt-1}, a_{jt-1}, i_{jt-1}, p_{jt-1})\big) \\ &\quad + \xi_{jt} + \eta_{jt}.\end{aligned} \tag{57}$$

Unfortunately, this equation cannot identify β_k and β_a since it requires us to condition on a nonparametric function of $(k_{jt-1}, i_{jt-1}, a_{jt-1})$. More formally, since $k_{jt} = (1-\delta)k_{jt-1} + i_{jt-1}$ (and $a_{jt} = a_{jt-1} + 1$), there is no remaining independent variance in (k_{jt}, a_{jt}) to be used to identify β_k and β_a.

To avoid this problem, we need an explicit ability to solve for or estimate μ_{jt-1}. This would generally require demand side data. For example, the Berry, Levinsohn and Pakes (1995) demand estimation procedure produces estimates of a set of "unobserved product characteristics" which might be used as the μ_{jt}'s. Of course, once one brings in the demand side, there is other information that can often be brought to bear on the problem. For example, the production function estimates should imply estimates of marginal cost which, together with the demand system, would actually determine prices in a "static" Nash pricing equilibrium (see the first section of this chapter). We do not pursue this further here.

Finally we move to the case where there are two sources of productivity growth, one evolving as a controlled Markov process, and one as an exogenous Markov process. In this case the production function is written as

$$y_{jt} = \beta_0 + \beta_k k_{jt} + \beta_a a_{jt} + \beta_l l_{jt} + \omega^1_{jt} + \omega^2_{jt} + \eta_{jt},$$

where ω^1_{jt} is the controlled, and ω^2_{jt} is the exogenous, first order Markov process.

Assume now that we have data on both R&D expenditures, say r_t, which is the input of the controlled process, and a "technology indicator" or T_t (like patents, or licensing fees) which is an output of the controlled process. As above, assume the policy functions for physical and R&D investment are a bijection, so we can write

$$\begin{aligned}\omega^1_{jt} &= \Upsilon^{-1}_{1t}(k_{jt}, a_{jt}, i_{jt}, r_{jt}),\\ \omega^2_{jt} &= \Upsilon^{-1}_{2t}(k_{jt}, a_{jt}, i_{jt}, r_{jt}).\end{aligned} \tag{58}$$

Now the first stage consists of using the technology indicator to isolate ω^1_{jt}. In other words, we assume

$$T_{jt} = \omega^1_{jt}\theta + \eta_{2jt}, \tag{59}$$

where $\eta_{2,t}$ is mean independent of all the controls. We then substitute a nonparametric function of $(k_{jt}, a_{jt}, i_{jt}, r_{jt})$ for ω^1_{jt} in Equation (59). This provides us with an estimate of $\omega^1_{jt}\theta$, say $\widehat{\Upsilon}^{-1}_{1tj}$.

Our second stage mimics the first stage of OP except we treat $\widehat{\Upsilon}^{-1}_{1tj}$ as an input. That is, we estimate

$$y_{jt} = \beta_l l_{jt} + \phi(k_{jt}, a_{jt}, i_{jt}, r_{jt}) + \eta_{jt}, \tag{60}$$

where

$$\phi(k_{jt}, a_{jt}, i_{jt}, r_{jt}) = \beta_k k_{jt} + \beta_a a_{jt} + \theta^{-1}\widehat{\Upsilon}^{-1}_{1jt} + \omega^2_{jt}.$$

Then, without a selection correction, the third stage becomes

$$\tilde{y}_{jt} = \beta_k k_{jt} + \beta_a a_{jt} + \tilde{g}\big(\phi_{jt-1} - \beta_k k_{jt-1} - \beta_a a_{jt-1}, \widehat{\Upsilon}^{-1}_{1jt-1}\big) + \xi_{jt} + \eta_{jt}.$$

Once again, we can modify this to allow for selection by using the propensity score as an additional determinant of $\tilde{g}(\cdot)$.

Buettner (2004b) explores a related extension to OP. While he only allows one unobserved state variable, he does allow the distribution of ω_{jt} to evolve endogenously over time, i.e. firms invest in R&D and these investments affect the distribution of ω_{jt} (conditional on ω_{jt-1}).[48] Unlike the above, Buettner does not assume that a "technology indicator" is observed. He develops a dynamic model with investments in R&D and physical capital that generates invertible policy functions such that the first stage of OP can be directly applied (and the labor coefficient can be estimated). However, second stage estimation is problematic, since the conditional expectation of ω_{jt} now depends on the full state vector through the choice of R&D. Furthermore, with the endogenous productivity process, he cannot rely on exogenous variation (such as changes in the economic environment over time) for identification. It remains to be seen whether this problem can be solved.

Greenstreet (2005) proposes and utilizes an alternative model/methodology that, while related to the above procedures, does not require the first stage inversion. This is a very nice attribute since as a result, the procedure does not rely at all on the key scalar unobservable and monotonicity assumptions of the OP/LP/ACF procedures. Greenstreet achieves this by making a different assumption on firms' information sets. Specifically, instead of observing ω_{jt} and η_{jt} individually (after production at t), firms only ever observe the *sum* $\omega_{jt} + \eta_{jt}$. Because of this alternative informational assumption, the econometrician does not need the first-stage inversion to recreate the information set of the firms. While this does avoid the scalar unobservable and monotonicity assumptions, Greenstreet's approach still relies on similar timing assumptions, involves a slightly more complicated learning process than the above procedures (requiring Kalman filtering), and also generates some new initial conditions problems that require additional assumptions to solve.

2.5. Concluding remark

The increase in the availability of plant and/or firm level panels together with a desire to understand the efficiency implications of major environmental and policy changes has led to a renewed interest in productivity analysis. Most of this analysis is based on production function estimates, and the literature has found at least two empirical regularities. First, there are indeed large efficiency differences among firms and those differences are highly serially correlated. Second, at least in many environments, to obtain realistic production function estimates the researcher must account for the possibility of simultaneity and selection biases.

Put differently, to study either the changes in the allocative efficiency of production among firms of differing productivities, or the correlates of productivity growth within individual establishments, we first have to isolate the productivity variable itself. Since firms' responses to the changes in the environment being studied typically depend on

[48] Recall that "endogenous" evolution of ω_{jt} is problematic for IV approaches.

how those changes impacted their productivity, movements in productivity cannot be isolated from changes in input and exit choices without an explicit model of how those choices are made.

The appropriateness of different models of how these decisions are made will undoubtedly depend on the environment being studied. We have presented a number of alternatives, and discussed their properties. However this is an empirically driven subfield of estimation, and there are undoubtedly institutional settings where alternative frameworks might be better to use. It is not the precise framework that is important, but rather the fact that productivity studies must take explicit account of the fact that changes in productivity (or, if one prefers, sales for a given amount of inputs) in large part determine how firms respond to the changes being studied, and these must be taken into account in the estimation procedure.

3. Dynamic estimation

This chapter considers structural estimation of dynamic games. Despite a blossoming empirical literature on structural estimation of static equilibrium models, there has been relatively little empirical work to date on estimation of dynamic oligopoly problems. Four exceptions are Gowrisankaran and Town (1997), Benkard (2004), Jofre-Bonet and Pesendorfer (2003), and Ryan (2006). The literature's focus on static settings came about not because dynamics were thought to be unimportant to market outcomes, but rather because empirical analysis of dynamic games was seen as too difficult. In particular, while some of the parameters needed to analyze dynamic games could be estimated without imposing the dynamic equilibrium conditions, some could not and, until very recently, the only available methods for estimating these remaining parameters were extremely burdensome, in terms of both computation time and researcher time.

This computational complexity resulted from the need to compute the continuation values to the dynamic game in order to estimate the model. The direct way of obtaining continuation values was to compute them as the fixed point to a functional equation, a high order computational problem. Parameter values were inferred from observed behavior by computing the fixed point that determines continuation values at different trial parameter values, and then searching for the parameter value that makes the behavior implied by the continuation values "as close as possible" to the observed behavior. This "nested fixed point" algorithm is extremely computationally burdensome because the continuation values need to be computed many times.

However, a recent literature in industrial organization [Aguirregabiria and Mira (2007), Bajari, Benkard and Levin (2007), Jofre-Bonet and Pesendorfer (2003), Pakes, Ostrovsky and Berry (2007), and Pesendorfer and Schmidt-Dengler (2003)] has developed techniques that substantially reduce the computational and programming burdens of estimating dynamic games. This literature extends a basic idea that first appeared in the context of single agent problems in Hotz and Miller (1993). Hotz and Miller (1993) provided a set of assumptions under which one could obtain a nonparametric estimate

of continuation values without ever computing the fixed point.[49] Rust (1994) suggests the extension of these ideas to the context of dynamic games. The recent literature in industrial organization has shown that, at least under a certain set of assumptions, these approaches can be extended to estimate continuation values in a wide variety of dynamic games, even in the presence of multiple equilibria.

This chapter summarizes the currently available techniques for estimating dynamic games, concentrating on this recent literature. The chapter proceeds as follows. We first outline the goals of the estimation procedure and consider what might be gained by modelling dynamics in an oligopoly situation. Then we present a general framework for dynamic oligopoly problems, with three simple examples from the recent literature. Next we overview existing estimation methods, providing details for the three examples. We conclude with a brief discussion of techniques available to ameliorate one (of many) outstanding problems; that of serially correlated unobserved state variables.

We note that there are at least two issues that appear in the literature and are not considered here. First we do not consider identification issues (at least not directly). Our feeling is that many of the parameters determining behavior in dynamic games can be estimated without ever computing an equilibrium, and those parameters that remain depend on the nature of the problem and data availability. Second, we do not consider "timing" games, such as those in Einav (2003) and in Schmidt-Dengler (2003). Our only excuse here is our focus on the evolution of market structure in oligopolies.

3.1. Why are we interested?

One contribution of the recent literature is that it provides a means of obtaining information about certain parameters that could not be obtained via other methods. For example, the sunk costs of entry and the sell-off values (or costs) associated with exit are key determinants in the dynamics of market adjustments to policy and environmental changes. Knowledge of the level of sunk costs is critical, for example, in a regulatory authority's decision of whether to approve a merger, or in the analysis of the likely impacts of changes in pension policy on shut down decisions. However, actual data on sunk costs are extremely rare. Besides being proprietary, and thus hard to access, sunk costs can also be very difficult to measure. Thus, in many cases the only option for learning the extent of sunk costs may be to infer them from equilibrium behavior using other variables that we can observe. Since sunk costs are only paid once upon entry, while firms may continue to operate for many periods, inferring the level of sunk costs from equilibrium behavior requires a dynamic framework. Similar arguments can be made regarding the parameters determining, among other diverse phenomena, the transaction

[49] In related work Olley and Pakes (1996) use nonparametrics to get around the problem of computing the fixed point needed to obtain an agent's decision rule in a multiple agent framework; but they use the nonparametric estimates to control for unobservables and do not recover the implied estimates of continuation values.

costs of investments (including installment, delivery, and ordering costs), the costs of adjusting output rates or production mix, and the extent of learning-by-doing.

There are a number of other uses for techniques that enable us to empirically analyze dynamic games. For example, there are many industries in which an understanding of the nature of competition in prices (or quantities) requires a dynamic framework. In such cases, the empirical literature in industrial organization has often used static models to approximate behavior that the authors are well aware is inherently dynamic. For example, there has been much work on identifying and estimating the form of competition in markets [e.g. Bresnahan (1982, 1987), Lau (1982)]. This literature typically compares a static Nash equilibrium with particular static "collusive" pricing schemes. In reality, the set of collusive pricing schemes that could be supported in equilibrium depends on the nature of the dynamic interactions [e.g. Abreu, Pearce and Stacchetti (1986), Green and Porter (1984), Rotemberg and Saloner (1985), Fershtman and Pakes (2000)]. A related point is that static price or quantity setting models are known to be inappropriate when future costs depend directly on the quantity sold today, as in models with learning by doing or adjustment costs, and/or when future demand conditions depend on current quantities sold, as in models with durable goods, experience goods, the ability to hold inventory, and network externalities.

Similarly, most of the existing empirical literature on entry relies on two-period static models. While these models have proven very useful in organizing empirical facts, the two period game framework used makes little sense unless sunk costs are absent. Therefore, the results are not likely to be useful for the analysis of policy or environmental changes in a given market over time. This leaves us with an inability to analyze the dynamic implications of a host of policy issues, and there are many situations where dynamics may substantially alter the desirability of different policies. For example, Fershtman and Pakes (2000) show that because collusive behavior can help promote entry and investment, it can enhance consumer welfare. Similarly, a static analysis would typically suggest that mergers lower consumer welfare by increasing concentration, whereas a dynamic analysis might show that allowing mergers promotes entry, counterbalancing the static effects.

3.2. Framework

This section outlines a framework for dynamic competition between oligopolistic competitors that encompasses many (but not all) applications in industrial organization. Examples that fit into the general framework include entry and exit decisions, dynamic pricing (network effects, learning-by-doing, or durable goods), dynamic auction games, collusion, and investments in capital stock, advertising, or research and development. The defining feature of the framework is that actions taken in a given period affect future payoffs, and future strategic interaction, by influencing only a set of *commonly observed* state variables. In particular, we will assume that all agents have the same information to use in making their decisions, up to a set of disturbances that have only transitory effects on payoffs.

We use a discrete time infinite horizon model, so time is indexed by $t = 1, 2, \ldots, \infty$. At time t, prevailing conditions are summarized by a state, $\mathbf{s}_t \in S \subset \mathbb{R}^G$, that reflects aspects of the world relevant to the payoffs of the agents. Relevant state variables might include firms' production capacities, the characteristics of the products they produce, their technological progress up to time t, the current market shares, stocks of consumer loyalty, or simply the set of firms that are incumbent in the market. We assume that these state variables are commonly observed by the firms. Note that we have not yet specified which state variables are observed by the econometrician. This distinction will be made in the applications below.

Given the state $\mathbf{s}_t$ at date t, the firms simultaneously choose actions. Depending on the application, the firms' actions could include decisions about whether to enter or exit the market, investment or advertising levels, or choices about prices and quantities. Let $a_{it} \in A_i$ denote firm i's action at date t, and $\mathbf{a}_t = (a_{1t}, \ldots, a_{Nt})$ the vector of time t actions, where N_t is the number of incumbents in period t (entry and exit, and hence N_t, are endogenous in these models).

We also assume that before choosing its action each firm, i, observes a *private* shock $\nu_{it} \in \mathbb{R}$, drawn independently (both over time and across agents) from a distribution $G(\cdot|\mathbf{s}_t)$.[50] Private information might derive from variability in marginal costs of production that result, say, from machine breakdowns, or from the need for plant maintenance, or from variability in sunk costs of entry or exit. We let the vector of private shocks be $\boldsymbol{\nu}_t = (\nu_{1t}, \ldots, \nu_{Nt})$.

In each period, each firm earns profits equal to $\pi_i(\mathbf{a}_t, \mathbf{s}_t, \nu_{it})$. Profits might include variable profits as well as any fixed or sunk costs, including the sunk cost of entry and the selloff value of the firm. Conditional on the current state, $\mathbf{s}_0$, and the current value of the firm's private shock, ν_{i0}, each firm is interested in maximizing its expected discounted sum of profits

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \pi_i(\mathbf{a}_t, \mathbf{s}_t, \nu_{it}) \middle| \mathbf{s}_0, \nu_{i0}\right], \tag{61}$$

where the expectation is taken over rival firms' actions in the current period as well as the future values of all state variables, the future values of the private shock, and all rivals' future actions. We assume firms have a common discount factor β.

The final aspect of the model is to specify the transitions between states. We assume that the state at date $t+1$, denoted $\mathbf{s}_{t+1}$, is drawn from a probability distribution $P(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$. The dependence of $P(\cdot|\mathbf{s}_t, \mathbf{a}_t)$ on the current period actions $\mathbf{a}_t$ reflects the fact that some time t decisions may affect future payoffs, as is clearly the case if the relevant decision being modelled is an entry/exit decision or a long-term investment. Of course, not all the state variables necessarily depend on past actions; for example,

[50] Here we assume that firm i's private shock is a single scalar variable. However, as will be seen in the examples below, there is no conceptual difficulty in allowing the shock to be multi-dimensional.

one component of the state could be a transitory i.i.d. shock that affects only the current payoffs, such as an i.i.d. shock to market demand.

Note that we have assumed that firms' private information does not influence state transitions directly (i.e. it only influences transitions through its impact on a_{it}). For example, incumbent firms care only about whether or not a potential entrant enters the market, and not what the entrant's sunk cost of entry was. On the other hand this assumption does rule out applications where firms' investment outcomes are their private information [e.g. Fershtman and Pakes (2005)].

We are interested in equilibrium behavior. Because the firms interact repeatedly and the horizon is infinite, there are likely to be many Nash, and even subgame perfect equilibria, possibly involving complex behavioral rules. For this reason, we focus on pure strategy Markov perfect equilibria (MPE).

In our context a Markov strategy for firm i describes the firm's behavior at time t as a function of the commonly observed state variables and firm i's private information at time t. Formally, it is a map, $\sigma_i : S \times \mathbb{R} \to A_i$. A profile of Markov strategies is a vector, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$, where $\boldsymbol{\sigma} : S \times \mathbb{R}^n \to A$. A Markov strategy profile, $\boldsymbol{\sigma}$, is a MPE if there is no firm, i, and alternative Markov strategy, σ_i', such that firm i prefers the strategy σ_i' to the strategy σ_i given its opponents use the strategy profile $\boldsymbol{\sigma}_{-i}$. That is, $\boldsymbol{\sigma}$ is a MPE if for all firms, i, all states, $\mathbf{s}$, and all Markov strategies, σ_i',

$$V_i(\mathbf{s}, \nu_i | \sigma_i, \boldsymbol{\sigma}_{-i}) \geqslant V_i\big(\mathbf{s}, \nu_i \big| \sigma_i', \boldsymbol{\sigma}_{-i}\big). \tag{62}$$

If behavior is given by a Markov profile $\boldsymbol{\sigma}$, firm i's present discounted profits can be written in recursive form

$$V_i(\mathbf{s}, \nu_i | \boldsymbol{\sigma}) = \mathbb{E}_{\boldsymbol{\nu}_{-i}}\Bigg[\pi_i\big(\boldsymbol{\sigma}(\mathbf{s}, \boldsymbol{\nu}), \mathbf{s}, \nu_i\big) + \beta \int V_i\big(\mathbf{s}', \nu_i' \big| \boldsymbol{\sigma}\big)\, \mathrm{d}G\big(\nu_i' \mid \mathbf{s}'\big)\, \mathrm{d}P\big(\mathbf{s}' \big| \boldsymbol{\sigma}(\mathbf{s}, \boldsymbol{\nu}), \mathbf{s}\big)\Bigg]. \tag{63}$$

3.2.1. Some preliminaries

The framework above is a generalization of the Ericson and Pakes (1995) model. The existence proofs for that model that are available have incorporated additional assumptions to those listed above [see Gowrisankaran (1995), and Doraszelski and Satterthwaite (2007)]. Typically, however, the algorithms available for computing an equilibrium do find an equilibrium even when the available sets of sufficient conditions for existence are not satisfied (i.e. the algorithm outputs policies and values that satisfy the fixed point conditions that define the equilibrium up to a precision determined by the researcher). There may, however, be more than one set of equilibrium policies [for an explicit example see Doraszelski and Satterthwaite (2007)].

If the regularity conditions given in Ericson and Pakes (1995) are satisfied, each equilibrium generates a finite state Markov chain for the $\mathbf{s}_t$ process. That is, the vector of state variables can only take on a finite set of values, a set we will designate by $\mathcal{S}$, and

the distribution of the future $\{\mathbf{s}_\tau\}_{\tau=t}^{\infty}$ conditional on all past history depends only on the current value of $\mathbf{s}_t$. Every sequence from this finite state Markov chain will, in finite time, wander into a subset of the states called a recurrent class or an $\mathcal{R} \subset \mathcal{S}$, and once in $\mathcal{R}$ will stay there forever. Every $s \in \mathcal{R}$ will be visited infinitely often.[51]

Throughout we assume that agents' perceptions of the likely future states of their competitors depend only on $\mathbf{s}_t$ (i.e. we assume that $\mathbf{s}_t$ is a complete description of the state variables observed by the firms). As detailed by Pakes, Ostrovsky and Berry (2007), this implies that there is only one equilibrium policy for each agent that is consistent with the data generating process; at least for all $\mathbf{s}_t \in \mathcal{R}$. To see this it suffices to note that since we visit each point in $\mathcal{R}$ infinitely often, we will be able to consistently estimate the distribution of future states of each firm's competitors given any $\mathbf{s}_t \in \mathcal{R}$. Given that distribution, each agent's best response problem is a single agent problem. Put differently, since reaction functions are generically unique, once the agent knows the distribution of its competitors' actions, its optimal policy is well defined. Thus, given the data generating process, policies are well defined functions of the parameters and the state variables. Consequently, standard estimation algorithms can be used to recover them.[52]

Finally, in all of the examples below we will assume that the discount factor, β, is one of the parameters that is known to the econometrician. It is a straightforward extension to estimate the discount parameter. However, our focus here is on obtaining estimates of parameters that we have little other information on.

3.2.2. Examples

The framework above is general enough to cover a wide variety of economic models. We provide three examples below. In general the objects that need to be recovered in the estimation are the period profit function, $\pi(\cdot)$, the transition probabilities, $P(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$, and the distribution of the private shocks, $G(\cdot|\mathbf{s})$.

EXAMPLE 1 *(A simple model of entry/exit).* This example is based on Pakes, Ostrovsky and Berry (2007). Let the state variables of the model be given by a pair, $\mathbf{s}_t = (n_t, \mathbf{z}_t)$, where n_t denotes the number of firms active at the beginning of each period, and $\mathbf{z}_t$ is a vector of profit shifters that evolve exogenously as a finite state. In the model, operating profits are determined solely by these variables. In any period, t, in which a firm is active it earns profits equal to

$$\tilde{\pi}(n_t, \mathbf{z}_t; \theta).$$

[51] Formally, the dynamics of the model are described by a Markov matrix. Each row of the matrix provides the probability of transiting from a given $\mathbf{s}$ to each possible value of $\mathbf{s} \in \mathcal{S}$. Ericson and Pakes (1995) also provide conditions that imply that the Markov matrix is ergodic, that is there is only one possible $\mathcal{R}$.

[52] Note that if our data consists of a panel of markets, this implicitly assumes that, conditional on s_t, the policy rule (our σ) in one market is the same as in the other.

The model focuses on entry and exit. In each period, each incumbent firm receives a random draw, denoted ϕ_{it}, determining the selloff value of the firm. The selloff values are assumed to be private information. However, their distribution is commonly known to the agents. The firm chooses to exit if the selloff value of the firm is greater than the expected discounted value of continuing in the market. Otherwise, the firm continues in the market.

Entry is described similarly. For ease of exposition, we assume that there are E potential entrants each period, where E is known to the agents.[53] Each period, each potential entrant firm receives a random draw, denoted κ_{it}, determining its sunk cost of entry. As above, the entry cost is private information, but its distribution is commonly known. The firm enters the market if the expected discounted value of entering is greater than the entry cost. Otherwise, the entrant stays out of the market and earns nothing.

To see how this model fits into the general framework, let $\chi_{it} = 1$ for any firm i that is active in the market in period t, and $\chi_{it} = 0$ otherwise. Note that we assume that when an incumbent firm exits, $\chi_{it} = 0$ thereafter. In that case the period profit function is

$$\pi_i(\mathbf{a}_t, \mathbf{s}_t, \nu_{it}) = \{\chi_{it} = 1\}\tilde{\pi}(n_t, \mathbf{z}_t; \theta) + (\chi_{it} - \chi_{i,t-1})^{-}\phi_{it} - (\chi_{it} - \chi_{i,t-1})^{+}\kappa_{it},$$

where the notation $\{\chi_{it} = 1\}$ denotes an indicator function that is one if the firm is active and zero otherwise, the notation $f^{+} \equiv \{f > 0\}f$, for any function f, and similarly $f^{-} \equiv \{f < 0\}|f|$. On the right-hand side, χ represents firms' actions, a; n and $\mathbf{z}$ represent the states, $\mathbf{s}$; and ϕ and κ represent the private shocks, ν.

Note that while this model does not allow for observed heterogeneity among incumbent firms, this can be achieved by allowing for multiple entry locations. We consider this extension below. Note further that this model is a special case of the Ericson and Pakes (1995) model in which investment is not modelled. We add investment back to the model in the next example.

EXAMPLE 2 *(An investment game with entry and exit).* This example is a straightforward extension of the Ericson and Pakes (1995) model due to Bajari, Benkard and Levin (2007). Similarly to the above example, there are a set of incumbent firms competing in a market. Firms are heterogeneous, with differences across firms described by their state variables, s_{it}, which are commonly known. For ease of exposition, we will omit any other exogenous profit shifters from the set of state variables.

Each period, firms choose investment levels, $I_{it} \geqslant 0$, so as to improve their state the next period. Investment outcomes are random, and each firm's investment affects only its own state so that there are no investment spillovers. Therefore, each firm's state variable, s_{it}, evolves according to a process $\Pr(s_{i,t+1}|s_{it}, I_{it})$.

Here are some examples of models that are consistent with this framework.

[53] The extension to a random number of entrants is straightforward. See Pakes, Ostrovsky and Berry (2007) for details.

(i) Firms' state variables could represent (one or more dimensions of) product quality, where investment stochastically improves product quality.
(ii) Firms' state variables could represent the fraction of consumers who are aware of the firm's product, where investment is a form of advertising that increases awareness [e.g. Doraszelski and Markovich (2007)].
(iii) Firms' state variables could represent capital stock, where investment increases a firm's capital stock.

Firms earn profits by competing in a spot market. Because quantity and price are assumed not to influence the evolution of the state variables, they are determined in static equilibrium conditional on the current state. In any period, t, in which a firm is active in the market it earns profits equal to

$$q_{it}(\mathbf{s}_t, \mathbf{p}_t; \theta_1)\big(p_{it} - \text{mc}(s_{it}, q_{it}; \theta_2)\big) - C(I_{it}, \nu_{it}; \theta_3), \tag{64}$$

where q_{it} is quantity produced by firm i in period t, $\mathbf{p}_t$ is the vector of prices, mc is the marginal cost of production, ν_{it} represents a private shock to the cost of investment, $\theta = (\theta_1, \theta_2, \theta_3)$ is a parameter vector to be estimated, and we have assumed that the spot market equilibrium is Nash in prices.

The model also allows for entry and exit. Each period, each incumbent firm has the option of exiting the market and receiving a scrap value, Φ, which is the same for all firms (this differs from the prior example in which there is a distribution of exit costs). There is also one potential entrant each period with a random entry cost, κ_{it}.[54] The entrant enters if the expected discounted value of entering exceeds the entry cost. As above, the entry cost is assumed to be private information, but its distribution is commonly known.

Relative to the general framework above, current period returns are given by

$$\begin{aligned}\pi_i(\mathbf{a}_t, \mathbf{s}_t, \nu_{it}) = \{\chi_{it} = 1\}\big[q_{it}(\mathbf{s}_t, \mathbf{p}_t; \theta_1)\big(p_{it} - \text{mc}(s_{it}, q_{it}; \theta_2)\big) - C(I_{it}, \nu_{it}; \theta_3)\big] \\ + (\chi_{it} - \chi_{i,t-1})^{-}\Phi - (\chi_{it} - \chi_{i,t-1})^{+}\kappa_{it}.\end{aligned}$$

On the right-hand side, prices (p), investment (I), and entry/exit (χ) are the actions (a), while the private shocks are the shock to investment (ν_{it}) and the entry cost (κ_{it}).

EXAMPLE 3 *(A repeated auction game with capacity constraints).* This example is based on Jofre-Bonet and Pesendorfer (2003). In this example, a set of incumbent contracting firms compete in monthly procurement auctions. The auctions are heterogeneous because the contracts that become available each month are of differing size and scope. The firms bidding on the contracts are also heterogeneous as each has a different cost of completing each contract. In a given month, each firm also has a different backlog of contracts, which might affect its ability to take on new contracts.

Let $\mathbf{z}_t$ be the characteristics of the contract to be auctioned in month t, including both the contract size (in dollars), and the number of months required to complete the

[54] It is straightforward to generalize the model to have a random number of potential entrants each period.

contract. We assume that $\mathbf{z}_t$ evolves exogenously as a finite state. Let $\omega_{i,t}$ be the backlog of work for firm i in period t and $\boldsymbol{\omega}_t = (\omega_{1,t}, \ldots, \omega_{N,t})$ be the vector of backlogs. A firm's backlog of work represents the remaining size in dollars, and the remaining number of days left until completion of each contract previously won by the firm. It therefore evolves deterministically depending on the current auction outcome according to the map

$$\boldsymbol{\omega}_{t+1} = \Gamma(\boldsymbol{\omega}_t, \mathbf{z}_t, j),$$

where j is the winner of the time t auction and the map Γ is known. The state variables of the model are $\mathbf{s}_t = (\boldsymbol{\omega}_t, \mathbf{z}_t)$. All states are assumed to be common knowledge.

Each firm also has a different cost, c_{it}, for each contract that is private information to the firm. Bidders' costs are drawn independently from a distribution $G(c_{it}|\omega_{it}, \omega_{-it}, \mathbf{z}_t)$ that is commonly known.

In each period, each firm views its cost for the contract being offered and then chooses a bid, b_{it}. Each firm earns current profits equal to

$$\pi_i(\mathbf{a}_t, \mathbf{s}_t, \nu_{it}) = (b_{it} - c_{it})\Big\{b_{it} \leqslant \min_j(b_{jt})\Big\}, \tag{65}$$

where the indicator function takes the value one if firm i submits the lowest bid and hence wins the auction (assume there are no ties). On the right-hand side the bids (b_{jt}) are the action variables (a_t) and the costs c_{it} are the private shocks (ν_{it}).

Note that the state variables do not directly enter current profits in this model. However, the state variables influence all firms' costs and hence a firm's optimal bid depends on the current state both through its own costs directly and through the firm's beliefs about the distribution of rivals' bids. For the same reason, expected profits are also a function of the current state.

Note also that an important distinction between the investment model above and this example is that, in this example, each firm's choice variable (in this case, its bid) affects the evolution of all firms' states. In the investment model above, a firm's investment affects only the evolution of its own state. This distinction is important because many I.O. models share this feature. For example, models of dynamic pricing (learning by doing, network effects, or durable goods) would have this feature when firms compete in prices (though not if firms compete in quantities). Such models can be placed in the general EP framework we have been using, but to do so we need to adjust that framework to allow the control that affects the distribution of current profits (bids, quantities, or prices) to also have an impact on distribution of future states; see the discussion in Pakes (1998). We note that to our knowledge Jofre-Bonet and Pesendorfer (2003) were the first to show that a two-step estimation approach was feasible in a dynamic game.

3.3. Alternative estimation approaches

In order to conduct policy analysis in any of the economic models above, it is typically necessary to know all of the parameters of the model, including the profit function, the

transition probabilities, and the distribution of the exogenous shocks. Often many of the parameters can be estimated "off line", that is, without needing to compute equilibria of the dynamic game. At one extreme here is Benkard's (2004) analysis of the commercial aircraft industry. He was able to obtain a large amount of cost data on sunk as well as marginal costs which, together with generally available information on demand, enabled him to estimate all the parameters he needed off line. Given these parameters he could focus on computing the dynamic implications of alternative policies.

However, such an approach is rarely possible. More typically, at least cost data are unavailable, either because they are proprietary and hence difficult for researchers to access, or because they are hard to measure. In static settings we often solve the problem of a lack of cost data by inferring marginal costs from their implications in an equilibrium pricing equation. A similar approach can be taken in this dynamic setting. However, characterizing the relationship between the data generating process and equilibrium play in the models above is complicated by the fact that the model involves repeated interactions.

Observed behavior in the model represents the solution to a maximization problem that involves both the profit function, which typically has a known parametric form, and the value function, which results from equilibrium play and therefore has unknown form. For example, the value of entering a market depends both on current profits, and expected future profits, which in turn depend on future entry and exit behavior. In order to describe the data generating process, then, we need the ability to compute the equilibrium continuation values.

Thus, conceptually, estimation of dynamic models can be separated into two main parts. The first part involves obtaining the continuation values for a given parameter value, θ. The second part is to use the continuation values obtained in the first part to maximize an objective function in the parameters, θ. Note that the continuation values must be obtained for many different values of θ in order to perform this maximization, and thus the first part is the source of most of the computational burden of the estimation. The key differences in estimation approaches lie in the details of how each of these two parts is performed.

3.3.1. The nested fixed point approach

The nested fixed point approach is a logical extension of the method of Rust (1987) to games. The general idea is as follows:

1. Given a parameter vector, θ, compute an equilibrium to the game, $V(\mathbf{s}; \theta)$, numerically, using the computer.
2. Use the computed values, $V(\mathbf{s}; \theta)$, to evaluate an objective function based on the sample data.
3. Nest steps one and two in a search routine that finds the value of θ that maximizes the objective function.

A framework capable of computing equilibria to models like those above has existed for some time [Pakes and McGuire (1994)], and recent papers by Pakes and McGuire

(2001), Doraszelski and Judd (2004), and Weintraub, Benkard and Van Roy (2007) enable significant improvements in computational times, at least in some problems [for a discussion of these, and other alternatives, see Doraszelski and Pakes (2007)]. All of these algorithms rely on similarities between the dynamic framework above and dynamic programming problems. The general idea of these algorithms is to start with an initial guess at the value function, $V^0(\mathbf{s};\theta)$, and substitute that into the right-hand side of the Bellman equation (Equation (63)). Then, at each state point and for each firm, solve the maximization equation on the right-hand side of (63) yielding a new estimate of the value function, $V^1(\mathbf{s};\theta)$. This procedure is iterated until convergence is achieved, so that the new and old value functions are the same. Unlike single agent problems, in the context of a game, convergence of the algorithm is not guaranteed (the mapping is not a contraction) and, indeed, initial iterations will often seem to move away from equilibrium. However, in practice the algorithms typically converge and, once they do, the value functions obtained must represent an equilibrium.

An important feature of the nested fixed point algorithm is that the first step is performed without using any data. As a result, the value functions are obtained precisely; that is, they contain no sampling error. This lack of sampling error makes the second part of the algorithm, in which the parameters are estimated, straightforward.

On the other hand the algorithm is computationally burdensome. For models rich enough to use in empirical work, it is often difficult to compute an equilibrium even once, and in the nested fixed point algorithm it is necessary to compute an equilibrium once for each iteration of the maximization routine; implying that up to hundreds, if not thousands, of fixed points must be calculated. Moreover, setting up an efficient algorithm often requires a large amount of complex computer programming, creating a relatively large burden on researcher time. As a result there are very few examples in the literature where the nested fixed point algorithm has been applied to estimate parameters.

One exception is Gowrisankaran and Town (1997), who use a nested fixed point approach to apply a model similar to the investment model above to data for the hospital industry. In each iteration of the estimation they compute an equilibrium using the algorithm of Pakes and McGuire (1994). They then estimate the model using a GMM objective function that matches cross-sectional moments such as average revenue per hospital, average expenditures per hospital, average investment per hospital, and average number of hospitals of each type (nonprofit and for-profit) per market. The nested fixed point approach was feasible in their application because their model was parsimonious and there were never more than three hospitals in any market in the data.

Another difficulty with the nested fixed point algorithm arises from the fact that dynamic oligopoly models can admit more than one equilibria. While the assumptions given above in principle allow the researcher to use the data to pick out the correct equilibrium, actually achieving this selection using the nested fixed point algorithm is likely to be difficult. Moreover, equilibrium selection has to take place for every candidate value of the parameters to be estimated. Alternative sets of assumptions could be used to select different equilibria, but unless we were willing to assume "a priori" that equi-

librium was unique, somehow we must investigate the issue of the relationship between the equilibrium computed in the algorithm, and that observed in the data.

3.3.2. Two-step approaches

The biggest obstacle to implementing the nested fixed point algorithm in practice is the heavy computational burden that results from the need to compute equilibria for each trial parameter value. Fortunately, the recent literature [Aguirregabiria and Mira (2007), Bajari, Benkard and Levin (2007), Jofre-Bonet and Pesendorfer (2003), Pakes, Ostrovsky and Berry (2007), and Pesendorfer and Schmidt-Dengler (2003)] has derived methods for estimating dynamic oligopoly models that impose the conditions of a dynamic equilibrium without requiring the ability to compute an equilibrium. The new literature sidesteps the equilibrium computation step by substituting nonparametric functions of the data for the continuation values in the game. These nonparametric estimates are in general much easier to compute than the fixed point calculations in the nested fixed point algorithm. As a result, these methods have substantially lower computational burden.

Below we outline five different two-step methods of estimating dynamic games. The overall approach is similar throughout, but each method does both the first and second steps of the estimation differently. To our knowledge, Hotz and Miller (1993) were the first to show that it was possible to estimate the continuation values in a dynamic programming problem nonparametrically instead of computing them. In a single agent dynamic discrete choice problem, Hotz and Miller showed that the agent's dynamic choice problem mimics a static discrete choice problem with the value functions replacing the mean utilities. Thus, the agent's continuation values can be obtained nonparametrically by first estimating the agent's choice probabilities at each state, and then inverting the choice problem to obtain the corresponding continuation values. This inversion is identical to the one commonly used in discrete choice demand estimation to obtain the mean utilities.

We begin our discussion of estimation by showing that if the game has only discrete actions, and there is one unobserved shock per action for each agent in the game, then under the information structure given in the general framework above, estimators very similar to those of Hotz and Miller (1993) can still be used [see also Aguirregabiria and Mira (2007)]. Sticking with the single agent framework, Hotz et al. (1994) use estimated probabilities to simulate sample paths. They then calculate the discounted value of utility along these paths, average those values for the paths emanating from a given state, and use these averages as the continuation values at that state. The Bajari, Benkard and Levin (2007) paper discussed below shows that related ideas can be used to incorporate continuous controls into estimation strategies for dynamic games.

Pakes, Ostrovsky and Berry (2007) also consider dynamic discrete games but, instead of inverting the choice problem, they estimate the continuation values directly by computing (approximately) the average of the discounted values of future net cash flows

that agents starting at a particular state actually earned in the data (at least up to the parameter vector of interest). Econometrically, they use a nonparametric estimate of the Markov transition probabilities that determine the evolution of the state of the system to form an analytic estimate of the probability weighted average of the discounted returns earned from different states. Given equilibrium play, these averages will converge to the true expected discounted value of future net cash flow, that is of the continuation values we are after.

Bajari, Benkard and Levin (2007) instead begin by projecting the observed actions on the state variables to compute nonparametric estimates of the policy functions of each agent at each state. Then they use the estimated policies to simulate out the discounted values of future net cash flows. This procedure is computationally light even in models with large state spaces and is easily applied to models with continuous controls, such as investment, quantity, or price (including models with both discrete and continuous controls like the investment game above). Given equilibrium play, the continuation values obtained in this fashion will be consistent estimates of the continuation values actually perceived by the agents.

Berry and Pakes (2002) provide an alternative approach for estimating models with continuous controls that is likely to be useful when the dynamic environment is complex, but sales and investment data are available. They assume that current period net returns are observable up to a parameter vector to be estimated, but do not require that the state variables of the model be observed, or even specified (so it would not be possible to estimate policy functions conditional on those state variables as in Bajari, Benkard and Levin). They derive an estimating equation from the first order condition for the continuous control (investment in our example) by substituting observed profit streams for the expected profit streams, and noting that the difference must be orthogonal to information known at the time investment decisions are made.

Jofre-Bonet and Pesendorfer (2003) provide an estimator for the dynamic auction model. They show that it is possible to derive an expression for the equilibrium continuation values in the auction game that involves only the bid distributions. Since bids are observed, the bid distributions can be recovered nonparametrically from the data and then substituted into these expressions. Provided that agents are bidding close to optimally, the continuation values obtained from this procedure will be consistent estimates of the continuation values perceived by the agents.

In many of the cases we consider several of the methods could be used in estimation. In these cases it is not currently known how the methods compare to one another on such dimensions as computational burden and econometric efficiency. Hybrid methods are also possible in which features of two or more of the approaches could be combined. We expect these issues to be sorted out in the future.

Finally, there are also some costs associated with the two-step approaches. First, because the continuation values are estimated rather than computed, they contain sampling error. This sampling error may be significant because these models often have state spaces that are large relative to the available data. As we will see below, this influences the properties of the second step estimators in important ways. To summarize, the

choice of second stage estimation method will be influenced as much or more by a need to minimize small sample bias caused by error in the continuation value estimates as it is by the desire to obtain asymptotic efficiency.

Second, for the two step approaches to produce estimators with desirable properties the data must visit a subset of the points repeatedly. Formally the requirement for the limit properties of the estimators is that all states in some recurrent class $\mathcal{R} \subseteq \mathcal{S}$ be visited infinitely often. Moreover, equilibrium strategies must be the same every time each point in $\mathcal{R}$ is visited. Whether or not this assumption is reasonable for the problem at hand depends on the nature of the available data and the institutional setting which generated it. If the data consists of a time series on one market then we would require stationarity of the process over time. There are different ways to fulfill this requirement in panels (i.e. when we follow a cross section of markets over time); one possibility is that the initial state in each market is a random draw from a long run ergodic distribution. Note that the nested fixed point approach has a weaker data requirement.

These costs must be weighed against the benefit that the two-step estimators eliminate most of the computational burden of the nested fixed point approach. Indeed, the entire two-step algorithm might well have less computational burden than one iteration of the nested fixed point algorithm.

3.4. A starting point: Hotz and Miller

Because of the similarity of this section to the previous literature on single agent problems, we will keep this section short, concentrating mainly on extending Hotz and Miller to games. For more detail on the approach in single agent problems see Hotz and Miller (1993), Hotz et al. (1994), Magnac and Thesmar (2002), and Rust (1994). See also Aguirregabiria and Mira (2007) and Pesendorfer and Schmidt-Dengler (2003) for a discussion in the context of entry games.

The idea behind Hotz and Miller's estimation method for single agent problems is to set up a dynamic discrete choice problem such that it resembles a standard static discrete choice problem, with value functions taking the place of standard utility functions. This allows a two step approach in which a discrete choice model is used as a first step for recovering the value functions, and the parameters of the profit function are recovered in a second step once the value functions are known.

We make two simplifying assumptions that will assist in the exposition. First, we suppose that agents' current profits do not depend on rivals' actions (though they do depend on rival's states whose evolution depends on those actions). Second, we assume that the unobserved shocks are additive to profits. In that case, current profits are given by

$$\pi_i(\mathbf{a}_t, \mathbf{s}_t, \nu_{it}) = \tilde{\pi}(a_{it}, \mathbf{s}_t) + \nu_{it}(a_{it}),$$

where ν_{it} is agent i's vector of profitability shocks and $\nu_{it}(a_{it})$ is the shock associated with agent i's action a_{it}.

The first assumption simplifies the agents' choice problem because, if agents' current profits depend on rivals' actions then, since rivals' actions depend on their own current shocks, in its own maximization problem each agent would have to integrate current profits over all rivals' current actions. This would not change the overall approach but would complicate the computations below [we would need to integrate over distributions of competitors outcomes to compute the expected profits; see Aguirregabiria and Mira (2007) for a model in which profits do depend on rivals' actions]. The second simplification, additive separability in the private shocks, is also not strictly required. If the error terms entered profits nonlinearly then we could rewrite the problem in terms of expected profits and an additively separable projection error and work with that framework instead. However, such an approach does have the unattractive property that it changes the interpretation of the profit function. Thus, it is typically the case that in practice people assume that the profit function has additive structural error terms.

With these assumptions the Bellman equation can be simplified to (suppressing the subscripts)

$$V(\mathbf{s}, \boldsymbol{\nu}) = \max_{a} \left\{ \tilde{\pi}(a, \mathbf{s}) + \boldsymbol{\nu}(a) + \beta \int V(\mathbf{s}', \boldsymbol{\nu}') \, \mathrm{d}G(\boldsymbol{\nu}'|\mathbf{s}') \, \mathrm{d}P(\mathbf{s}'|\mathbf{s}, a) \right\}. \tag{66}$$

Equation (66) represents a discrete choice problem in which the mean utilities are given by

$$v_a(\mathbf{s}) = \tilde{\pi}(a, \mathbf{s}) + \beta \int V(\mathbf{s}', \boldsymbol{\nu}') \, \mathrm{d}G(\boldsymbol{\nu}'|\mathbf{s}') \, \mathrm{d}P(\mathbf{s}'|\mathbf{s}, a). \tag{67}$$

Thus, since the private shocks are independent across time and across agents, the choice probabilities for a given agent can be generated in the usual manner of a static discrete choice problem

$$\Pr(a|\mathbf{s}) = \Pr\big(v_a(\mathbf{s}) + \boldsymbol{\nu}(a) \geqslant v_{a'}(\mathbf{s}) + \boldsymbol{\nu}(a'), \forall a'\big). \tag{68}$$

Assuming that the data consists of a large sample of observations on states and actions, the probability of each action at each state, $\Pr(a|\mathbf{s})$, can be recovered from the data. In that case, the left-hand side of (68) is known, at least asymptotically. Let $P(\mathbf{s})$ be the vector of choice probabilities for all feasible actions. Hotz and Miller show that for any distribution of the private shocks there is always a transformation of the choice probabilities such that

$$v_a(\mathbf{s}) - v_1(\mathbf{s}) = Q_a\big(\mathbf{s}, P(\mathbf{s})\big). \tag{69}$$

That is, the differences in the choice specific value functions can be written as a function of the current state and the vector of choice probabilities. The transformation on the right-hand side is the same as the inversion used in the discrete choice demand estimation literature. Berry (1994) proves that the solution is unique. Berry, Levinsohn and Pakes (1995) provide a transformation from the data to the mean utilities which is

a contraction, and hence enables the researcher to actually compute the mean utilities (for more discussion see the first part of this chapter).

In general, this transformation can be used to recover the normalized choice specific value functions, $v_a - v_1$, at each state, using the estimated choice probabilities. If the distribution of the private shocks is known, the mapping does not depend on any unknown parameters. For example, in the case of the logit

$$Q_a\big(\mathbf{s}, P(\mathbf{s})\big) = \ln\big(\Pr(a|\mathbf{s})\big) - \ln\big(\Pr(a_1|\mathbf{s})\big). \tag{70}$$

However, in general the mapping may be a function of unknown parameters of the distribution of the private shocks.

Note that, as in static discrete choice models, only the value differences can be recovered nonparametrically. Thus, some further information is required to obtain the values themselves. This difficulty is not just a feature of this particular estimation approach, but comes from the underlying structure of the discrete choice framework, in which only utility differences are identified from the observed choices. One consequence of this is that, even if the discount factor and the distribution of private shocks are completely known, the profit function cannot be recovered nonparametrically [see Magnac and Thesmar (2002) for a detailed proof and analysis for single agent dynamic discrete choice problems, and Pesendorfer and Schmidt-Dengler (2003) for results extended to dynamic discrete games]. This feature is inherent to the dynamic discrete choice setup and carries through to the context of a dynamic discrete game. As noted earlier our feeling is that the appropriate resolution of identification issues, such as this one, is context specific and will not be discussed here.

To obtain the continuation values from the choice specific values we can use the fact that

$$V(\mathbf{s}, \boldsymbol{\nu}) = \max_a \big\{ v_a(s) + \boldsymbol{\nu}(a) \big\}. \tag{71}$$

Because the continuation values are obtained by inverting from the observed choice probabilities, the structure of the profit function has not yet been imposed on them, and they are not yet functions of the profit function parameters. In order to estimate the profit function parameters, Hotz and Miller iterate the Bellman equation once, inserting the estimated continuation values on the right-hand side,

$$\begin{aligned} \widehat{V}(\mathbf{s}; \theta) = \int \max_a \Big\{ & \tilde{\pi}(a, \mathbf{s}; \theta) + \boldsymbol{\nu}(a) \\ & + \beta \int \widehat{V}(\mathbf{s}', \boldsymbol{\nu}') \, \mathrm{d}G(\boldsymbol{\nu}'|\mathbf{s}') \, \mathrm{d}P(\mathbf{s}'|\mathbf{s}, a) \Big\} \, \mathrm{d}G(\boldsymbol{\nu}|\mathbf{s}). \end{aligned} \tag{72}$$

Note that for some distributions such as those of type GEV the integral on the right-hand side has an analytic form. In other cases it can be simulated.

These new estimates of the continuation values contain the profit function parameters (θ) and can be used in an estimation algorithm to estimate θ. The way this is typically

done is to compute new predicted choice probabilities, (68), based on the new continuation value estimates, $\widehat{V}(\mathbf{s};\theta)$. Then, these choice probabilities can be used to construct either a pseudo-likelihood or some other GMM objective function that matches the model's predictions to the observed choices.

As noted above, the nonparametric estimates of the continuation values and transition probabilities on the right-hand side of (72) introduce estimation error into the second stage objective function nonlinearly. Hotz and Miller show that if this estimation error disappears quickly enough then the estimator obtained is consistent and asymptotically normal. However, there are other methods that may be preferable in this context to a pseudo-likelihood. Because of the nonlinearity of the pseudo-likelihood in the continuation values, estimation error in the continuation values causes increased small sample bias in the parameter estimates obtained using this method. We discuss methods that at least partially address this problem in the next section.

3.5. Dynamic discrete games: Entry and exit

In this section we consider estimation of the entry/exit game in example one using the methods of Pakes, Ostrovsky and Berry (2007) (henceforth POB). We maintain the assumption that all of the state variables, $(n_t, \mathbf{z}_t)$, are observed and that the number of entrants (e_t) and exitors (x_t) are also observed. Entry and exit costs are assumed not to be observed and are the objects of interest in the estimation. We discuss the possibilities for estimation when there are one or more unobserved state variables in Section 3.8.1.

Consider first exit behavior. Redefining the value function from the start of a period, prior to the point at which the private scrap value is observed, the Bellman equation for incumbent firms is given by (t subscript suppressed)

$$V(n,\mathbf{z};\theta) = \tilde{\pi}(n,\mathbf{z};\theta) + \beta\mathbb{E}_{\phi}\big[\max\big\{\phi_i, \mathrm{VC}(n,\mathbf{z};\theta)\big\}\big], \tag{73}$$

where VC denotes the continuation value of the firm, which equals

$$\mathrm{VC}(n,\mathbf{z};\theta) \equiv \sum_{\mathbf{z}',e,x} V(n+e-x,\mathbf{z}';\theta)P(e,x|n,\mathbf{z},\chi=1)P(\mathbf{z}'|\mathbf{z}). \tag{74}$$

In the above equation, e and x denote the number of entering and exiting firms, and $P(e,x|n,\mathbf{z},\chi=1)$ denotes the incumbent's beliefs about the likely number of entrants and exitors starting from state $(n,\mathbf{z})$ *conditional* on the incumbent itself continuing $(\chi=1)$.

If the equilibrium continuation values, $\mathrm{VC}(n,\mathbf{z};\theta)$, were known, then it would be straightforward to construct a likelihood function since the probability of exit is given by

$$\Pr(i \text{ exits}|n,\mathbf{z};\theta) = \Pr\big(\phi_i > \mathrm{VC}(n,\mathbf{z};\theta)\big), \tag{75}$$

and is independent across firms. Thus, we need to find a simple way to construct the equilibrium continuation values using observed play.

The continuation values represent the expected discounted value of future profits conditional on the incumbent continuing. They are a function of the profit function, $\tilde{\pi}(n, \mathbf{z}; \theta)$, which determines future profits at each state $(n, \mathbf{z})$, and the processes determining the evolution of the state variables, n and z. The profit function is known up to the parameters, θ. Therefore, in order to construct the continuation values as a function of the parameters, we need only estimate the evolution of the number of firms, which is determined by entry and exit, and the evolution of the profit shifters, $P(\mathbf{z}'|\mathbf{z})$. The easiest way to do this is to use their empirical counterparts. Starting from a certain state, to estimate the evolution of the number of firms we can use the actual evolution of the number of firms each time that state was observed in the data. Similarly, we can use the observed evolution of the profit shifters to estimate the process $P(\mathbf{z}'|\mathbf{z})$. That way the estimated continuation values reflect, approximately, the actual profits of firms that were observed in the data. The next subsection outlines this process in detail.

3.5.1. Step 1: Estimating continuation values

To facilitate estimation of the continuation values, it helps to rewrite the Bellman equation in terms of the continuation values, VC,

$$\begin{aligned}\mathrm{VC}(n, \mathbf{z}; \theta) = \sum_{n', z'} \big[& \tilde{\pi}(n', \mathbf{z}'; \theta) \\ & + \beta \mathbb{E}_{\phi} \big[\max \big\{ \phi_i, \mathrm{VC}(n', \mathbf{z}'; \theta) \big\} \big] \big] P(n'|n, \mathbf{z}, \chi = 1) P(\mathbf{z}'|\mathbf{z}), \end{aligned} \tag{76}$$

where to shorten the notation we let $n' \equiv n + e - x$.

Next, rewrite (76) in vector form. Let $\mathrm{VC}(\theta)$ be the $\#\mathcal{S} \times 1$ vector representing $\mathrm{VC}(n, \mathbf{z}; \theta)$ for every $(n, \mathbf{z})$ pair, and define $\tilde{\pi}(\theta)$ similarly. Also let M^i be the $\#\mathcal{S} \times \#\mathcal{S}$ matrix whose (i, j) element is given by $P(n_j|n_i, z_i, \chi = 1) P(z_j|z_i)$. This is the matrix whose rows give us the equilibrium transition probabilities from a particular $(n, \mathbf{z})$ to each other possible $(n, \mathbf{z})$. Note that if we were not conditioning on $\chi = 1$ an unbiased estimate of the rows of this matrix could be obtained by simply counting up the fraction of transits from $(n, \mathbf{z})$ that were to each other state. Since the continuation value the agent cares about is the continuation value should the agent continue, these estimates have to be modified for conditioning on $\chi = 1$, see the discussion below.

With this notation, (76) becomes

$$\mathrm{VC}(\theta) = M^i \tilde{\pi}(\theta) + \beta M^i \mathbb{E}_{\phi} \big[\max \big\{ \phi_i, \mathrm{VC}(\theta) \big\} \big]. \tag{77}$$

In this last equation, $\tilde{\pi}(\theta)$ is a known vector (up to θ). In a structural model the distribution of ϕ would also typically be known up to a parameter vector. Therefore, the only unknowns in the equation are M^i and $\mathrm{VC}(\theta)$. If M^i were known, $\mathrm{VC}(\theta)$ could be calculated as the solution to the set of equations (77). We discuss the estimation of M^i below and turn first to the solution for $\mathrm{VC}(\theta)$.

One of the insights of POB is that the expectations term on the right-hand side of (77) can sometimes be simplified, making computation of $\mathrm{VC}(\theta)$ simple. Expanding the

expectations term at a single state (n, z) gives

$$\begin{aligned}
&\mathbb{E}_\phi\big[\max\{\phi_i, \mathrm{VC}(n, \mathbf{z}; \theta)\}\big] \\
&\quad = \Pr\big(\phi_i < \mathrm{VC}(n, \mathbf{z}; \theta)\big) * \mathrm{VC}(n, \mathbf{z}; \theta) \\
&\qquad + \Pr\big(\phi_i > \mathrm{VC}(n, \mathbf{z}; \theta)\big) * \mathbb{E}_\phi\big[\phi_i \big| \phi_i > \mathrm{VC}(n, \mathbf{z}; \theta)\big] \\
&\quad = \big(1 - p_x(n, \mathbf{z})\big) * \mathrm{VC}(n, \mathbf{z}; \theta) + p_x(n, \mathbf{z}) * \mathbb{E}_\phi\big[\phi_i \big| \phi_i > \mathrm{VC}(n, \mathbf{z}; \theta)\big],
\end{aligned}$$

where $p_x(n, z)$ is the probability of exit at state (n, z). Provided that the distribution of scrap values is log-concave, the above equation is a contraction mapping [see Heckman and Honoré (1990)]. In that case, given estimates of M^i and p^x, the equation can be solved for $\mathrm{VC}(\cdot)$ in a straightforward manner. Moreover, when the distribution of scrap values is exponential, a distribution often thought to be reasonable on *a priori* grounds,

$$\mathbb{E}_\phi\big[\phi_i \big| \phi_i > \mathrm{VC}(n, \mathbf{z}; \theta)\big] = \sigma + \mathrm{VC}(n, \mathbf{z}; \theta),$$

where σ is the parameter of the exponential, and

$$\begin{aligned}
&\mathbb{E}_\phi\big[\max\{\phi_i, \mathrm{VC}(n, \mathbf{z}; \theta)\}\big] \\
&\quad = \big(1 - p_x(n, \mathbf{z})\big) * \mathrm{VC}(n, \mathbf{z}; \theta) + p_x(n, \mathbf{z}) * \big[\mathrm{VC}(n, \mathbf{z}; \theta) + \sigma\big] \\
&\quad = \mathrm{VC}(n, \mathbf{z}; \theta) + \sigma p_x(n, \mathbf{z}).
\end{aligned}$$

Substituting this expression into (77) and iterating gives

$$\begin{aligned}
\mathrm{VC}(\theta) &= M^i\big[\tilde{\pi}(\theta) + \beta\sigma p_x\big] + \big(M^i\big)^2\big[\tilde{\pi}(\theta) + \beta\sigma p_x\big] + \big(M^i\big)^3 \mathrm{VC}(\theta) + \cdots \\
&= \sum_{\tau=1}^{\infty} \big(M^i\big)^\tau \big[\tilde{\pi}(\theta) + \beta\sigma p_x\big] \\
&= \big(I - \beta M^i\big)^{-1} M^i\big[\tilde{\pi}(\theta) + \beta\sigma p_x\big]. \qquad (78)
\end{aligned}$$

The only thing that remains is to estimate M^i and p_x using the data. Both can be estimated in a variety of different ways, but the simplest approach, and the one supported by POB's Monte Carlo results, is to use their empirical counterparts. Let

$$T(n, \mathbf{z}) = \big\{t\colon\ (n_t, \mathbf{z}_t) = (n, \mathbf{z})\big\}$$

be the set of periods in the data with the same state $(n, \mathbf{z})$. Then, the empirical counterpart to p_x is

$$\hat{p}_x(n, \mathbf{z}) = \frac{1}{\#T(n, \mathbf{z})} \sum_{t \in T(n, \mathbf{z})} \frac{x_t}{n}.$$

Due to the Markov property, $\hat{p}_x(n, \mathbf{z})$ is a sum of independent draws on the exit probability, and therefore it converges to $p_x(n, \mathbf{z})$ provided $\#T(n, \mathbf{z}) \to \infty$.

Similarly, the matrix M^i can be estimated element-by-element using

$$\widehat{M}^i_{i,j} = \frac{\sum_{t \in T(n_i, z_i)} (n_i - x_t) 1\{(n_{t+1}, \mathbf{z}_{t+1}) = (n_j, z_j)\}}{\sum_{t \in T(n_i, z_i)} (n_i - x_t)}.$$

This expression weights the actual observed transitions from (n_i, z_i) in different periods by the number of incumbents who actually continue in those periods. This weighting corrects the estimated transition probabilities for the fact that incumbents compute continuation values under the assumption that they will continue in the market.

Note that because this procedure uses empirical transition probabilities it never requires continuation values or transition probabilities from points not observed in the data. As a result there is no need to impute transition probabilities or continuation values for states not visited.[55] Since typical data sets will only contain a small fraction of the points in $\mathcal{S}$, this reduces computational burden significantly.

Substituting the estimated transition and exit probabilities into (78) provides an expression for the estimated continuation values

$$\widehat{\mathrm{VC}}(\theta, \sigma) = \left(I - \beta \widehat{M}^i\right)^{-1} \widehat{M}^i \left[\tilde{\pi}(\theta) + \beta\sigma \hat{p}_x\right]. \tag{79}$$

Note first that the estimates of continuation values using the expression in (79) are, approximately, the averages of the discounted values of the incumbents who did continue.[56] This makes the relationship between the data and the model transparent. Provided only that the specification of the profit function is correct, the actual average of realized continuation values should be close to the expected continuation values used by the agents in making their decisions.

Second, note how easy it is to compute the estimated continuation values. If the discount factor is known, then,

$$\widehat{\mathrm{VC}}(\theta, \sigma) = \widetilde{A}\tilde{\pi}(\theta) + \tilde{a}\sigma, \tag{80}$$

where $\widetilde{A} = (I - \beta\widehat{M}^i)^{-1}\widehat{M}^i$ and $\tilde{a} = \beta(I - \beta\widehat{M}^i)^{-1}\hat{p}_x$. Both $\widetilde{A}$ and $\tilde{a}$ are independent of the parameter vector and can therefore be computed once and then held fixed in the second step of the estimation.

Finally, note that the parameters of the entry distribution do not enter into the calculation of the continuation values. The reason for this is that sunk costs of entry are paid only once at the time of entry. After that, the sunk costs distribution only affects profits indirectly through rival firms' entry decisions. Thus, all that matters for computing continuation values is the probability of entry, not the associated level of sunk costs. As a result the computational burden of the model does not depend in any major way on the form of the entry cost distribution, a fact which is particularly useful when we consider models with multiple entry locations below.

Entry behavior can be described similarly. A potential entrant enters the market if the expected discounted value of entering is greater than the entry cost, i.e. if χ^e is the

[55] Strictly speaking this is only true if the last period's state in the data was visited before. If it were not we would have to impute transition probabilities for it.

[56] This is only approximately true because the transitions for all firms that reached a state $(n, \mathbf{z})$ are used to compute transitions for each firm, so information is pooled across firms in computing the continuation values.

indicator function which is one if the potential entrant enters and zero elsewhere

$$\beta \,\mathrm{VE}(n, \mathbf{z}; \theta) \geqslant \kappa,$$

where

$$\mathrm{VE}(n, \mathbf{z}; \theta) \equiv \sum_{\mathbf{z}', e, x} V(n + e - x, \mathbf{z}'; \theta) P\big(e, x \big| n, \mathbf{z}, \chi^e = 1\big) P(z' | \mathbf{z}),$$

similarly to VC before. The main difference here is that the entrant is not active in the current period and therefore forms beliefs slightly differently from the incumbent.

The incumbent forms beliefs conditional on it remaining active. The entrant forms beliefs based on it *becoming* active. In vector form, the expression for the entrants' continuation values is

$$\mathrm{VE}(\theta, \sigma) = M^e\big(\tilde{\pi} + \beta \,\mathrm{VC}(\theta) + \beta p_x \sigma\big),$$

where the elements of M^e represent a potential entrant's beliefs about the distribution over tomorrow's states conditional on that entrant becoming active. An estimator for M^e that is analogous to the one above is given by

$$\widehat{M}^e_{i,j} = \frac{\sum_{t \in T(n_i, z_i)} e_t 1\{(n_{t+1}, \mathbf{z}_{t+1}) = (n_j, z_j)\}}{\sum_{t \in T(n_i, z_i)} e_t}.$$

Accordingly, a consistent estimator of $\widehat{\mathrm{VE}}(\theta, \sigma)$ is given by

$$\widehat{\mathrm{VE}}(\theta, \sigma) = \widetilde{B} \tilde{\pi}(\theta) + \tilde{b} \sigma, \tag{81}$$

where $\widetilde{B} = \widehat{M}^e (I + \beta \widetilde{A})$, and $\tilde{b} = \beta \widehat{M}^e (\tilde{a} + \hat{p}_x)$.

3.5.2. Step 2: Estimating the structural parameters

If the continuation values (VE and VC) were known, any of a number of method of moments algorithms would provide consistent estimators of (θ, σ) and maximum likelihood would provide the efficient estimator. Since $\widehat{\mathrm{VE}}$ and $\widehat{\mathrm{VC}}$ are consistent estimators of the unknown continuation values, an obvious way to obtain a consistent estimator is to substitute them for VC and VE in any of these algorithms and proceed from there. For example, the implied "pseudo" maximum likelihood estimator would maximize

$$\begin{aligned} l(x_t, e_t | \theta, \sigma) = {} & (n_t - x_t) \log F^{\phi}\big[\widehat{\mathrm{VC}}(n_t, \mathbf{z}_t; \theta, \sigma)\big] \\ & + x_t \log\big[1 - F^{\phi}\big(\widehat{\mathrm{VC}}(n_t, \mathbf{z}_t; \theta, \sigma)\big)\big] \\ & + e_t \log F^{\kappa}\big[\widehat{\mathrm{VE}}(n_t, \mathbf{z}_t; \theta, \sigma)\big] \\ & + (E - e_t) \log\big[1 - F^{\kappa}\big(\widehat{\mathrm{VE}}(n_t, \mathbf{z}_t; \theta, \sigma)\big)\big], \end{aligned}$$

where F^{ϕ} is the distribution of scrap values and F^{κ} is the distribution of entry costs.

POB stress the importance of remembering that $\widehat{\text{VE}}$ and $\widehat{\text{VC}}$ contain sampling error. Though this sampling error does converge to zero with sample size, the fact that we have to estimate separate continuation values for each sample point means that, for standard sample sizes, the sampling error should not be ignored. This has implications both for the choice of estimators, and for how we compute standard errors for any given choice.

In this context there are two problems with the pseudo maximum likelihood estimator. First since it does not "recognize" that there is sampling error in the probabilities it uses, events can occur that the likelihood assigns zero probability to, no matter the value of (θ, σ) (even though the true probabilities of these events are nonzero; POB shows that this tends to occur in their two location model). If this happens even once in the data, the pseudo maximum likelihood estimator is not defined. Second, even if the pseudo-likelihood is well defined, its first order condition involves a function that is both highly nonlinear in, and highly sensitive to, the sampling error. The nonlinearity implies that the impact of the sampling error on the first order conditions will not average out over sample points. The sensitivity is seen by noting that the first order effect of the sampling error on the log likelihood will be determined by one over the probabilities of entry and exit, and these probabilities are typically quite small.

POB consider two alternatives to the likelihood approach. The first is a pseudo minimum χ^2 estimation algorithm that minimizes the sum of squares in the difference between the observed and predicted state specific entry and exit rates (i.e. the entry and exit rates for each observed $(n, \mathbf{z})$ pair), where the predicted state specific entry and exit rates are given by

$$E[x_t|n_t, \mathbf{z}_t] = n_t * \Pr\big(\phi_i > \text{VC}(n_t, \mathbf{z}_t; \theta, \sigma)\big), \quad \text{and}$$
$$E[e_t|n_t, \mathbf{z}_t] = E * \Pr\big(\kappa < \text{VE}(n_t, \mathbf{z}_t; \theta, \sigma)\big).$$

Their second estimator matches the overall entry and exit rates (across all observed state pairs) to those predicted by the model, or more generally takes a sum of squares in the differences between the predicted and actual entry and exit rates at different states multiplied by a *known* function of the state variables at those states.

They show that in finite samples the pseudo minimum χ^2 estimator has an extra bias term that reflects the sampling covariance between the estimated probability and its derivative with respect to the parameter vector, and their Monte Carlo evidence indicates that the extra bias term can have large effects. Thus they prefer the simplest method of moments algorithm and show that with moderately sized samples this estimator is both easy to calculate and performs quite well.

The second general point is that the variance of the second stage estimates, $(\hat{\theta}, \hat{\sigma})$, depends on the variance of the first stage estimates.[57] It is possible to use standard semiparametric formulae to obtain the asymptotic variance of the parameter estimates

[57] This follows from the fact that the derivative of the objective function with respect to the estimates of VC and VE are not conditional mean zero.

analytically. However these formula are somewhat complex and can be difficult to evaluate. Moreover, there is little reason to do the calculation. Since we have a complete model and the computational burden of obtaining estimates is minimal it is relatively easy to obtain estimates of standard errors from a parametric bootstrap.

For an empirical example which uses these techniques see Dunne et al. (2006). They estimate the parameters of a dynamic entry game from data on entry and exit of dentists and chiropractors in small towns. They first estimate the variable profit function (which depends on the number of active competitors) from observed data on revenues and costs. They then employ POB's method to provide estimates of the sunk costs of entry and of exit values. Their parameters could be used, for example, to predict the effect of a government subsidy intended to increase the number of medical service professionals in small towns.

3.5.3. Multiple entry locations

We now show how to generalize the model to allow for observed heterogeneity among incumbents. We do this by allowing entrants to choose from multiple entry locations. For ease of exposition, we will consider only two locations. However, expanding this to a larger number is straightforward.

Entrants have entry costs (κ_1, κ_2) in the first and second locations respectively, where entry costs are drawn from a distribution, $F^{\kappa}(\kappa_1, \kappa_2|\theta)$, that is independent over time and across agents. Note that we place no restrictions on F^{κ} so that entry costs of the same potential entrant at the different locations may be freely correlated. Once in a particular location, the entrant cannot switch locations, but can exit to receive an exit fee. Exit fees are an i.i.d. draw from the distribution $F_1^{\phi}(\cdot|\theta)$ if the incumbent is in location one, and an i.i.d. draw from $F_2^{\phi}(\cdot|\theta)$ if the incumbent is in location two.

The Bellman equation for an incumbent in the two location model is

$$V_1(n_1, n_2, \mathbf{z}; \theta) = \tilde{\pi}_1(n_1, n_2, \mathbf{z}; \theta) + \beta \mathbb{E}_{\phi}\big[\max\big\{\phi_i, \mathrm{VC}_1(n_1, n_2, \mathbf{z}; \theta)\big\}\big],$$

where the subscript "1" indicates the value function for a firm in location one and the continuation values are

$$\begin{aligned}\mathrm{VC}_1(n_1, n_2, \mathbf{z}; \theta) \equiv \sum_{\mathbf{z}', e_1, e_2, x_1, x_2} & V_1(n_1 + e_1 - x_1, n_2 + e_2 - x_2, \mathbf{z}'; \theta) \\ & \times P(e_1, e_2, x_1, x_2|n_1, n_2, \mathbf{z}, \chi = 1) P(\mathbf{z}'|\mathbf{z}).\end{aligned}$$

Behavior of incumbent firms is identical to before, with the probability of exit given by (75) except using the new continuation values. However, because they have potentially different continuation values and different scrap values, firms in location one will in general behave differently than firms in location two.

Behavior of potential entrant firms is different from before because potential entrant firms now have three options. They can enter location one, enter location two, or not enter at all. A potential entrant will enter into location 1 if and only if it is a better

alternative than both not entering anywhere, and entering into location 2, i.e. if

$$\beta \, \mathrm{VE}_1(n_1, n_2, \mathbf{z}; \theta) \geqslant \kappa_1 \quad \text{and}$$
$$\beta \, \mathrm{VE}_1(n_1, n_2, \mathbf{z}; \theta) - \kappa_1 \geqslant \beta \, \mathrm{VE}_2(n_1, n_2, \mathbf{z}; \theta) - \kappa_2.$$

The entry process therefore generates a multinomial distribution with probabilities derived from the inequalities above.

Firms' beliefs are now comprised of the probability of exit for an incumbent in location one, the probability of exit for an incumbent in location two, the three entry probabilities (location one, location two, or not at all), and the distribution determining the evolution of the profit shifters. In computing the continuation values we now have to consider all of these together.

Consistent estimates of a location-one incumbent's perceived transition probabilities from state (n_{i1}, n_{i2}, z_i) to (n_{j1}, n_{j2}, z_j) are obtained analogously to before using

$$\widehat{M}^{i,1}_{i,j} = \frac{\sum_{t \in T(n_{i1}, n_{i2}, z_i)} (n_{i1} - x_{t1}) 1\{(n_{t+1,1}, n_{t+1,2}, \mathbf{z}_{t+1}) = (n_{j1}, n_{j2}, z_j)\}}{\sum_{t \in T(n_{i1}, n_{i2}, z_i)} (n_{i1} - x_{t1})}.$$

Similarly, estimates of a potential location-one entrant's perceived transition probabilities can be obtained using

$$\widehat{M}^{e,1}_{i,j} = \frac{1}{\#T(n_{i1}, n_{i2}, z_i)} \times \frac{\sum_{t \in T(n_{i1}, n_{i2}, z_i)} e_{t1} 1\{(n_{t+1,1}, n_{t+1,2}, \mathbf{z}_{t+1}) = (n_{j1}, n_{j2}, z_j)\}}{\sum_{t \in T(n_{i1}, n_{i2}, z_i)} e_{t1}}.$$

As before these estimates are not exactly equal to the empirical frequency of state transitions but are a weighted average based on the fact that, when computing continuation values, an incumbent assumes it will continue, and a potential entrant assumes that it will enter.

As in the single location model, given the matrix inversion formula for continuation values, the computational burden of obtaining estimates for the parameters of the model is minimal. Indeed in their Monte Carlo results POB show that in two location models with relatively large data sets (on the order of 7500 observations) one finds estimates in under fifteen minutes on a five year old desktop computer. Most of that computation time is devoted to computing the Markov transition matrix and its inverse. The time required to compute the inverse can grow polynomially in the number of distinct states and, at least given market size, this typically increases with the number of locations. Whether it does or not depends on the structure of the matrix being inverted, and the way one computes the inverse. Models which only allow transitions to "near by" states, which are likely to dominate in I.O. applications, should not be as problematic in this respect.

Second, though the estimators remain consistent when the number of entry states is increased, their small sample properties may change. In particular, the estimates of the continuation values will become noisier in small samples and this is likely to cause

increased small sample bias and variance in the second stage estimates. POB show that the use of smoothing techniques, such as those discussed in the next section, can be helpful in this context.

3.6. Models with discrete and continuous controls: Investment games

In this section we consider Bajari, Benkard and Levin's (2007) (henceforth BBL) estimation approach in the context of the investment model in example two. The main conceptual difference in BBL's general approach that separates it from the above methods is that, instead of estimating continuation values directly, BBL first estimate policy functions. Then, the estimated policy functions are used to simulate the continuation values. As noted earlier this is similar to the single agent approach of Hotz et al. (1994), but BBL show that there are assumptions and techniques that allow the researcher to use this approach in a wide class of models, including models with both discrete and continuous controls such as investment models, some models of dynamic pricing, and dynamic auction problems. The assumptions used do carry with them some restrictions, and we will try to be clear on those restrictions below.

The presence of both discrete and continuous controls in the investment model affects both the first and second stage of the estimation. In particular, the second stage is augmented in order to incorporate information from firms' investment, as well as its entry/exit, choices. Additionally, when the stress is on investment we generally consider models with larger state spaces, and, as noted above, both computation of the estimates of continuation values, and the precision of those estimates, can become problematic. BBL introduce simulation techniques that, depending on the structure of the model, can cause a significant reduction in the computational burden of obtaining estimates of the continuation values. They also use techniques that smooth estimated continuation values across states to lower the mean square error of those estimates.

In the investment model from example two there are three policies (entry, exit, and investment) that are set in dynamic equilibrium, and one policy (price) that is set in static equilibrium. Since the pricing equilibrium is consistent with a large past literature on demand estimation, we will not consider estimation of the demand and marginal cost functions (θ_1 and θ_2) here as they would typically be estimated using existing methods. Instead, we will treat those parameters as if they were known and focus on estimation of the investment cost function (θ_3) and the entry and exit costs parameters.

We assume that all of the state variables, $\mathbf{s}$, are observed, as well as entry, exit, quantity, price, and investment levels. Entry and exit costs, the cost of investment function, and marginal costs are not observed. Note that it would be possible for some of the state variables to be unobserved as long as they could be recovered beforehand during estimation of the demand and cost systems. We discuss the issue of unobserved states further in Section 3.8.1.

Let $\tilde{\pi}_i(\mathbf{s})$ represent the profits earned by firm i in the spot market equilibrium at state $\mathbf{s}$. Since the demand and marginal cost functions are assumed to be known, the

function $\tilde{\pi}_i(\cdot)$ is also known, as the spot market equilibrium can be computed from these primitives.

Firms maximize the expected discounted value of profits. From the beginning of a period (prior to realization of the private shock), and for incumbent firms, this is

$$\mathbb{E}\sum_{t=0}^{\infty}\beta^t\big[\{\chi_{it}=1\}\big(\tilde{\pi}_i(\mathbf{s}_t)-C(I_{it},\nu_{it};\theta_3)\big)+(\chi_{it}-\chi_{i,t-1})^{-}\Phi\big|\mathbf{s}_0\big], \tag{82}$$

where $\chi_{it}=1$ indicates that the incumbent continues in the market at period t and $\chi_{it}=0$ indicates that the incumbent exits, and it is understood that each exiting firm receives the same exit value and never operates thereafter. Note that unlike in the entry/exit example above, in this model we assume that the incumbent's choice of its discrete control (whether or not to exit) is not subject to a random cost shock.

For expositional (and computational) simplicity we will assume the following quadratic cost of investment function

$$C(I,\nu;\theta_3)=\{I\geqslant 0\}\big(\theta_{3,0}+\theta_{3,1}I+\theta_{3,2}I^2+\theta_{3,3}\nu I\big). \tag{83}$$

The indicator function for $I\geqslant 0$ above allows for an adjustment cost that is incurred only if investment is nonzero. Zero investment is a phenomenon that is often observed and can easily result from either flatness of the value function reflecting low returns to investment [see Ericson and Pakes (1995)], or nonconvex investment costs [e.g. Caballero and Engel (1999)].

Potential entrant firms' expected discounted values are similar to (82), except that in the initial period they must pay a random entry cost, κ_{it}, in order to enter. We assume that entrants take one period to setup the firm and therefore do not earn profits in the spot market and do not invest until the subsequent period. For ease of exposition, we also assume that entrants always enter at the same initial state.[58]

3.6.1. Step 1: Estimating continuation values

The goal of the first step of the estimation procedure is to compute the continuation values given by the expected discounted values in (82), under equilibrium strategies. These expected discounted values are functions of the profits earned at each state and the probability distributions determining future states and actions conditional on the starting state, $\mathbf{s}_0$.

BBL compute estimates of the continuation values by first estimating policies for each state, then using the estimated policies to simulate sample paths of industry states and actions, and then evaluating discounted profits on each sample path. In order to do this we need both the ability to simulate sample paths of states and actions, and the ability to evaluate profits along those paths given the states and actions in each period.

[58] It is straightforward to allow entrants' initial state to be randomly determined.

Evaluating profits requires a (possibly flexible) description of the profit function and knowledge of the distribution of the private shocks, at least up to a parameter vector to be estimated. We treat these two as "primitives" of the dynamic model.

The evolution of the states depends on firms' entry, exit, and investment policies. BBL recover these policy functions using the observed data. In our investment model, private information is known to the firms before any actions are taken, so in MPE, strategies for investment, exit and entry are functions of both the states and this private information

$$I(\mathbf{s}_t, \nu_{it}), \quad \chi(\mathbf{s}_t, \nu_{it}), \quad \text{and} \quad \chi^e(\mathbf{s}_t, \kappa_{it}),$$

where χ is the exit policy function for incumbent firms, χ^e is the entry policy function and $\chi^e = 1$ indicates that the entrant enters the market. Since potential entrants cannot invest in the first period, entry strategies depend only on the random entry cost. Both the investment and exit strategies depend only on the shock to the marginal cost of investment.

Consider first exit and entry. The optimal exit policy has the form of a stopping rule

$$\chi_{i,t} = 1 \quad \text{iff} \quad \nu_{it} \leqslant \bar{\nu}(\mathbf{s}_t).$$

All we require is a nonparametric estimate of the probability that $\chi = 1$ conditional on $\mathbf{s}_t$. Similarly, there a critical entry level of κ conditional on $\mathbf{s}_t$ that determines entry, and the entry policy is obtained as a nonparametric estimate of the probability of entry conditional on $\mathbf{s}_t$. In both cases we also have the restriction that the policies are exchangeable in rivals' states. In models with large state spaces, such that there are some states in the data with few or zero observations, it would typically be optimal to employ some smoothing in these estimates. In their Monte Carlo studies, BBL found that local linear methods worked well for this.

As far as investment is concerned, one can show that, conditional on a firm continuing in the market, investment is a (weakly) monotone function of ν_{it}, $I(\mathbf{s}_t, \nu_{it})$. Thus, if we knew the distribution of investment at each state, $F(I_{it}|\mathbf{s}_t)$, we could map the quantiles of ν into investment levels at each state. More precisely, the investment policy function is given by

$$F^{-1}_{I|\mathbf{s}}\big(G(\nu|\mathbf{s})\big).$$

The function G is a primitive of the model known up to a parameter vector, and the function F can be recovered nonparametrically.

There is an additional complication and that is that investment is not observed for firms that exit the market, which happens if $\nu_{it} > \bar{\nu}(\mathbf{s}_t)$. However, since both the exit and investment rules are monotonic in the shock, this is handled easily. Conditional on a firm continuing in the market, we observe the distribution of investment conditional on $\mathbf{s}$ that corresponds to $\nu_{it} \leqslant \bar{\nu}(\mathbf{s}_t)$. Therefore, if we first estimate the probability of exit at each state, and then recover the distribution of investment at each state conditional on staying in the market, then we have a complete description of the optimal exit and

investment policy functions.[59] Note also that in the simulations below it is important that we maintain this link between exit and investment since one draw on the private shock to investment, ν_{it}, determines both policies.

If there was a second unobservable, say a random exit fee ϕ_{it}, then the exit decision would depend on both (ν_{it}, ϕ_{it}). The probability of exit could still be obtained as above, but the distribution of investment conditional on not exiting would depend on both ν_{it} and ϕ_{it}. Then, without further restrictions it would not be possible to invert the observed distribution of investment to obtain the policy decision as a function of ν conditional on $\mathbf{s}$ and not exiting.[60]

There also remains the question of how best to estimate the investment function, and this depends to some extent on its likely properties. Here it is important to keep in mind that investment is a complicated function of the primitives. Indeed the only restriction we have on its form is that it is exchangeable in the states of the competitors (which is already embodied in the definition of $\mathbf{s}$). Standard nonparametric approaches assume a certain amount of smoothness that is not necessarily guaranteed by the primitives of the model. The theoretical properties of the investment function in the EP model depend on the underlying properties of the family $\{P(s_{it+1}|I_{it}, s_{it})\}$. If conditional on s_{it} the points of support of this family do not depend on I_{it},[61] then by appropriate choice of primitives one can ensure that the investment function is smooth; see EP, the Monte Carlo evidence in BBL, and the generalizations of this in Doraszelski and Satterthwaite (2007). In their Monte Carlo studies BBL also found that local linear regression worked well for estimating the investment function.

Assume now that, for each state, we have consistent estimators of the entry probability, the exit probability, the investment distribution, and the distribution of future states. This is all one needs to compute consistent estimates of the continuation values in (82) as a function of the parameters. To do so analytically, however, would involve high dimensional integration, so what BBL do is show how to extend the "forward simulation" idea in Hotz et al. (1994) to simplify the analysis of the more complex problems they deal with.

Starting from a given state, $\mathbf{s}_0$, one draw is taken on the shock to the marginal cost of investment for the firm of interest, ν_i. This draw determines the firm's investment and exit policies through the estimated policy functions above (i.e. the same draw determines the correct quantile for both investment and exit, as discussed above). These policies, along with the state and the value of the private shock, determine current profits

[59] We cannot use the data to learn what an exiting firm would have invested had it stayed in the market, but it is not necessary to know this.

[60] Note that the problem here is that there is more than one error influencing the choice of investment. Therefore, a feasible alternative would be to allow a random exit cost but no shock to the marginal cost of investment.

[61] This assumption allows for stochastic outcomes to investment processes which is an assumption often made in Industrial Organization. However it does rule out the deterministic accumulation models traditionally used in growth theory.

as a function of the parameters. Draws are then taken for the investment shocks for the remaining incumbent firms, and one draw on the entry distribution is taken for the potential entrant. These draws, along with draws determining the outcomes of each firm's investment process, determine $\mathbf{s}_1$. The process is repeated to obtain one simulated path of states and the associated discounted stream of profits. Many such paths can be simulated to obtain an estimate of $V_i(\mathbf{s}_0)$. Consistency of the estimation algorithm requires that the number of simulated paths goes to infinity.

This forward simulation procedure is not too computationally burdensome, though one does have to hold one set of simulation draws in memory and use these same draws to evaluate the continuation values at the different values of θ tried in the estimation algorithm. Moreover, much of what computational burden remains disappears when we deal with models that are linear in the parameters. For example, suppose we consider the investment model above where the private shock to investment has a normal distribution, $\nu \sim N(0, 1)$. (The investment shock is normalized to be standard normal without loss of generality because its mean and variance parameters are absorbed into the parameters $\theta_{3,0}$ and $\theta_{3,3}$.) Since all of the parameters enter the continuation values linearly, they can be factored out as follows

$$\begin{aligned} V_i(\mathbf{s}_0; \sigma_i, \boldsymbol{\sigma}_{-i}) &= \mathbb{E}\sum_{t=0}^{\infty} \beta^t \{\chi_{it} = 1\} \tilde{\pi}_i(\mathbf{s}_t) - \theta_{3,0}\mathbb{E}\sum_{t=0}^{\infty} \beta^t \{\chi_{it} = 1\}\{I_{it} \geqslant 0\} \\ &\quad - \theta_{3,1}\mathbb{E}\sum_{t=0}^{\infty} \beta^t \{\chi_{it} = 1\}\{I_{it} \geqslant 0\} I_{it} \\ &\quad - \theta_{3,2}\mathbb{E}\sum_{t=0}^{\infty} \beta^t \{\chi_{it} = 1\}\{I_{it} \geqslant 0\} I_{it}^2 \\ &\quad - \theta_{3,3}\mathbb{E}\sum_{t=0}^{\infty} \beta^t \{\chi_{it} = 1\}\{I_{it} \geqslant 0\} I_{it} \nu_{it} \\ &\quad + \Psi\mathbb{E}\sum_{t=0}^{\infty} \beta^t (\chi_{it} - \chi_{i,t-1})^- \\ &\equiv \mathbf{W}_i\big(\mathbf{s}_0; \sigma_i, \boldsymbol{\sigma}_i'\big)' \begin{pmatrix} 1 \\ \theta_3 \\ \Psi \end{pmatrix}, \end{aligned} \tag{84}$$

where $W_i(\mathbf{s}_0, \sigma_i, \boldsymbol{\sigma}_{-i})$ represents the expected discounted value terms above when i follows policy σ_i and rival firms follow policy $\boldsymbol{\sigma}_{-i}$. The estimated continuation values are then computed by plugging in the estimated policies and simulating the expectations terms

$$\widehat{V}_i\big(\mathbf{s}_0; \hat{\sigma}_i, \hat{\boldsymbol{\sigma}}_i'\big) = \widehat{W}_i\big(\mathbf{s}_0; \hat{\sigma}_i, \hat{\boldsymbol{\sigma}}_i'\big)' \begin{pmatrix} 1 \\ \theta_3 \\ \Psi \end{pmatrix}.$$

The key observation here is that if the model is linear in the parameters, then the parameters factor out of the continuation value calculations. In that case the W terms need only be computed once, and the continuation values at different values of the parameter vector can be obtained by multiplying two small dimensional vectors.

This simplification is an extension of the one used in the entry/exit example above, except here we exploit linearity in the investment cost parameters as well as the linearity in the period profits. Since the continuation values need to be calculated many times in the second step of the estimation, and since computing continuation values is the primary source of computational burden, such simplifications can lead to a substantial reduction in the overall computational burden of the estimator.

3.6.2. Step 2: Estimating the structural parameters

As with the entry model above, once the continuation values have been estimated there are potentially many ways of estimating the structural parameters. The main difference is that now there is one continuous control variable (investment) in addition to the two discrete controls (entry/exit), and we want to use the information in the continuous control to help estimate the parameters. Accordingly all the issues that arose in the discussion of estimation of the entry/exit model are also relevant here. In particular, there is error in the estimated continuation values that can contaminate the second stage estimates, so it is desirable to find a second step estimator that is close to linear in the estimated continuation values.

There are at least three possible estimators: (i) an inequality estimator that finds a value of the parameter vector that insures that the observed policies generate higher simulated continuation values than alternative policies (see below), (ii) a method of moments estimator that fits the mean of the policies implied by the simulated continuation values (i.e. at each state in the data you substitute the simulated continuation values into the right-hand side of the Bellman equation (63) and solve for the optimal policy) to nonparametric estimates of the policies at each state, and (iii) a method of moments estimator that fits the nonparametric estimates of the distribution of the polices to the distribution of policies predicted by the simulated continuation values at each state. BBL provide Monte Carlo evidence on the first two of these. Here we review the inequality estimator, that is the estimator found by satisfying the optimality inequalities (62) that define the MPE for the simulated values.

At the true values of the parameters, for all states, $\mathbf{s}_0$, all firms, i, and all alternatives, σ_i', it must be that

$$\mathbf{W}_i(\mathbf{s}_0; \sigma_i, \boldsymbol{\sigma}_{-i})' \begin{pmatrix} 1 \\ \theta_3 \\ \Psi \end{pmatrix} \geqslant \mathbf{W}_i(\mathbf{s}_0; \sigma_i', \boldsymbol{\sigma}_{-i})' \begin{pmatrix} 1 \\ \theta_3 \\ \Psi \end{pmatrix}.$$

Let x refer to a particular $(i, \mathbf{s}_0, \sigma')$ combination, such that x indexes inequalities, and let

$$g(x; \theta_3, \Psi) = \big(\mathbf{W}_i(\mathbf{s}_0; \sigma_i, \sigma_{-i}) - \mathbf{W}_i\big(\mathbf{s}_0; \sigma_i', \sigma_{-i}\big)\big)' \begin{pmatrix} 1 \\ \theta_3 \\ \Psi \end{pmatrix}.$$

Then it must be the case that $g(x; \theta_3, \Psi) \geqslant 0$ at the true values of the parameters for every x.

A natural thing to do in estimation would be to compute g at the estimated policies from the first stage and then find the values of the parameters that best satisfy the entire set of inequalities. However, when there are continuous controls this is difficult because there are too many possible alternative policies. Instead, BBL use simulation to choose a small subset of the inequalities to impose in estimation. The inequalities can be chosen according to any random rule that selects all of them asymptotically. However, it is important to remember that the exact rule used will influence efficiency. In their Monte Carlo studies, for investment alternatives BBL use policies of the form

$$I'(\mathbf{s}_t, \nu_{it}) = \widehat{I}(\mathbf{s}_t, \nu_{it}) + \epsilon',$$

where ϵ' is drawn from a normal distribution with mean zero and standard deviation chosen by the researcher. Alternative entry and exit policies were chosen similarly by shifting the cutoff rule by an amount ϵ' drawn from a normal distribution.

Suppose n_i inequalities are sampled, and let $\hat{g}_{n_s}(x; \theta_3, \Psi)$ be a simulator for $g(x; \theta_3, \Psi)$ evaluated at the estimated policy functions, where n_s is the number of simulation draws used to simulate each $\mathbf{W}_i$ term. Then the inequality estimator minimizes the objective function

$$\frac{1}{n_I} \sum_{k=1}^{n_I} 1\big\{\hat{g}_{n_s}(x_k; \theta_3, \Psi) < 0\big\} \hat{g}_{n_s}(x_k; \theta_3, \Psi)^2.$$

Because the estimator is computationally light, it is easy to choose (n_I, n_s) to be large enough that the simulation contributes nothing to the variance of the estimator. All of the variance comes from error in the estimation of the continuation values. BBL work out the asymptotic distribution of the estimator. However, the expression is difficult to evaluate and in practice the simplest way to compute standard errors is to use subsampling or a bootstrap.

The inequality estimator has several advantages. One is that it is very easy to implement even in complex models. It is conceptually simple and requires a minimum of computer programming, the main programming burden being the forward simulation routine. Additionally, the method can be used with very little alteration even if the model is only set-identified. In that case, all that is required is to use an alternative method for computing standard errors [see BBL, as well as Chernozhukov, Hong and Tamer (2007) for details].

However, one potential disadvantage of the approach is that, similarly to the pseudo-likelihood methods shown in the examples above, the estimator is nonlinear in the first

stage estimates, and therefore the estimates obtained are likely to contain small sample bias. For that reason, BBL also tested a natural alternative estimator based on a set of moment conditions that match the observed choice data. The general idea of this estimator is to substitute the estimated continuation values into the right-hand side of the Bellman equation and then solve for an optimal policy rule conditional on those continuation values. This estimator is linear in the estimated continuation values, though those values are still nonlinear functions of the estimated policies. The expected value of the optimal policy is then matched against the average policy observed at each state in the data. In their Monte Carlo studies BBL found that this second estimator did help reduce small sample bias in the second stage estimates.

For an empirical example that uses these techniques see Ryan (2006). He estimates the parameters of a dynamic oligopoly model of US cement producers. In the first stage he estimates the static profits demand and cost parameters using demand data and a static equilibrium assumption. He also estimates the entry, exit, and investment policy functions using data on the set of firms operating in a panel of markets and their capacities. In the second stage he uses BBL's inequality estimator to estimate the sunk costs of entry and exit, as well as the adjustment costs of investment. He finds that the 1990 amendments to the Clean Air Act significantly raised the sunk costs of entry in the cement industry, and that a static analysis would have missed an associated welfare penalty to consumers.

3.6.3. An alternative approach

Berry and Pakes (2002) provide an alternative approach for estimating models with continuous choice variables that uses quite different assumptions from POB or BBL. They assume that profits are observable up to a parameter vector to be estimated, but do not require that the state variables that determine current and expected future profits are observed, and do not even require the researcher to specify what those state variables are. In applications where the environment is complex, but sales and investment data are quite good, this alternative set of assumptions can be quite attractive.

Let the random variable τ_i refer to the period in which firm i exits the market. Then, firm i's continuation value in the investment game starting at state $\mathbf{s}_0$ is

$$V_i(\mathbf{s}_t) = \mathbb{E}\left[\sum_{r=t}^{\tau_i} \beta^{r-t}\left(\tilde{\pi}_i(\mathbf{s}_r) - C\left(\sigma(\mathbf{s}_r);\theta_3\right)\right) + \beta^{\tau_i - t}\Phi \,\middle|\, \mathbf{s}_t\right], \tag{85}$$

where σ is the equilibrium policy function. Note that we have assumed there is no private shock to investment; an assumption that is needed for the consistency of this estimator.

Berry and Pakes note that, if firms have rational expectations, then the actual discounted stream of profits earned by a given firm is an unbiased estimate of its expected discounted profits. Suppose that profits ($\tilde{\pi}_{it}$), investment (I_{it}), and exit (χ_{it}) are observed. Then the actual discounted sum of profits earned by the firm (corresponding

to (85)) is

$$\widehat{V}_i(\mathbf{s}_t;\theta_3,\Phi) \equiv \sum_{r=t}^{\tau_i} \beta^{r-t}\big(\tilde{\pi}_{ir} - C(I_{ir};\theta_3)\big) + \beta^{\tau_i - t}\Phi, \tag{86}$$

where, in a slight abuse of notation, τ_i now refers to the actual period in which the firm exited. By rational expectations we have that, at the true values of the parameters,

$$\widehat{V}_i(\mathbf{s}_t;\theta_3,\Phi) = V_i(\mathbf{s}_t) + \epsilon_{it},$$

where $E[\epsilon_{it}|\mathbf{s}_t] = 0$.

A unique feature of the Berry and Pakes approach is that the estimated continuation values here are unbiased. However, in contrast to POB and BBL, Berry and Pakes (2002) do not have a first stage that provides consistent estimates of continuation values. Since the state variables are assumed not to be observed, there is no longer any way of identifying a set of data points that correspond to the same state vector. Thus, there is no way to average out across observations so as to obtain consistent estimates of the continuation values, as in POB and BBL.

Berry and Pakes get around the problem of having only unbiased, and not consistent estimates of continuation values, by using an estimating equation that has the error in the estimated continuation value entering linearly. More precisely, their estimating equation is derived from the first order condition for the firm's continuous control. Conditional on investment being strictly positive (a condition that is determined by the information available when the investment decision is made, and hence that is independent of the realization of ϵ_{it}), that first order condition is obtained by setting the derivative of (85) equal to zero. Using the cost of investment function after eliminating the i.i.d. shock to investment this gives us

$$\begin{aligned}
0 &= -\theta_{3,1} - 2*\theta_{3,2}*I_{it} + \beta\sum_{\mathbf{s}_{t+1}} V_i(\mathbf{s}_{t+1})\frac{\partial}{\partial I}P(\mathbf{s}_{t+1}|I_{it},\mathbf{s}_t,\chi_{it}=1)\\
&= -\theta_{3,1} - 2*\theta_{3,2}*I_{it}\\
&\quad + \beta\sum_{\mathbf{s}_{t+1}} V_i(\mathbf{s}_{t+1})\frac{\frac{\partial}{\partial I}P(\mathbf{s}_{t+1}|I_{it},\mathbf{s}_t,\chi_{it}=1)}{P(\mathbf{s}_{t+1}|I_{it},\mathbf{s}_t,\chi_{it}=1)}P(\mathbf{s}_{t+1}|I_{it},\mathbf{s}_t,\chi_{it}=1)\\
&= -\theta_{3,1} - 2*\theta_{3,2}*I_{it}\\
&\quad + \beta E\left[V_i(\mathbf{s}_{t+1})\frac{\partial \ln P(\mathbf{s}_{t+1}|I_{it},\mathbf{s}_t,\chi_{it}=1)}{\partial I}\middle|\mathbf{s}_t, I_{it}, \chi_{it}=1\right].
\end{aligned} \tag{87}$$

Adding and subtracting a term in $\widehat{V}(\mathbf{s}_{t+1};\theta_3,\Psi)$ gives

$$\begin{aligned}
0 &= -\theta_{3,1} - 2*\theta_{3,2}*I_{it} + \widehat{V}_i(\mathbf{s}_{t+1};\theta_3,\Phi)\frac{\partial \ln P(\mathbf{s}_{t+1}|I_{it},\mathbf{s}_t,\chi_{it}=1)}{\partial I}\\
&\quad + \eta_{it}(\theta_3,\Phi),
\end{aligned} \tag{88}$$

where we have defined

$$\eta_{it}(\theta_3, \Phi) \equiv \beta E\left[V_i(\mathbf{s}_{t+1}) \frac{\partial \ln P(\mathbf{s}_{t+1}|I_{it}, \mathbf{s}_t, \chi_{it} = 1)}{\partial I} \middle| \mathbf{s}_t, I_{it}, \chi_{it} = 1 \right] - \widehat{V}_i(\mathbf{s}_{t+1}; \theta_3, \Phi) \frac{\partial \ln P(\mathbf{s}_{t+1}|I_{it}, \mathbf{s}_t, \chi_{it} = 1)}{\partial I}, \tag{89}$$

and consequently

$$E\left[\eta_{it}(\theta_3, \Phi)\middle|\mathbf{s}_t\right] = 0, \tag{90}$$

at the true values of the parameters vector. Condition (90) follows from the twin facts that $V_i(\mathbf{s}_{t+1}) = \widehat{V}_i(\mathbf{s}_{t+1}; \theta_3, \Phi) - \epsilon_{i,t+1}$ and

$$E\left[\frac{\partial \ln P(\mathbf{s}_{t+1}|I_{it}, \mathbf{s}_t, \chi_{it} = 1)}{\partial I} \right] \epsilon_{i,t+1} = 0,$$

as the derivative is a function of information known at t. It follows that (88) provides a set of conditional moment restrictions that can be used as the basis for estimation.

There are a number of disadvantages of this approach. One that can potentially be corrected is that as presented in Berry and Pakes (2002) the algorithm does not incorporate the additional information in the data that comes from the choice of discrete controls (e.g. entry and exit), or from controls chosen to be at a corner of the choice set (e.g. $I_{i,t} = 0$). One could add a set of inequality constraints to the Berry–Pakes model to account for entry and exit and the $I_{i,t} = 0$ case. Also, as mentioned above, it is difficult to incorporate a shock to the cost of investment into this model.

However the major difference between this model and the other models discussed above is that Berry and Pakes do not need to specify and control for all the state variables in the dynamic system. This is an obvious advantage for complex problems. Of course, if we cannot identify and control for all the state variables of the system, we cannot make use of averaging techniques that enable us to use information on similar states to construct estimates of the policies and returns at a given state. In problems where the state variables are easy to identify and control for, averaging techniques can be very helpful in reducing variance. It remains to be seen if hybrids can be developed that make effective use of all of these techniques.

3.7. A dynamic auction game

In this section we consider estimation of the auction model in example three. This section closely follows Jofre-Bonet and Pesendorfer (2003) (henceforth JP). We assume that all bids, contract characteristics, and bidders' state variables are observed. A unique feature of the auction model is that the period payoff function is not a function of any unknown parameters. The goal of estimation, then, is to recover the distribution of bidders' privately known costs at each state.

Since the outcome of the auction affects not only current profits but also the firm's backlog, firms choose their bids so as to maximize the expected discounted value of future profits. Recall that $\mathbf{z}_t$ provides the characteristics of the contracts to be auctioned

in month t and evolves as a Markov process, $\omega_{i,t}$ provides the backlog of work of firm i in period t, and if $\boldsymbol{\omega}_t = (\omega_{1,t}, \ldots, \omega_{N,t})$, then $\omega_{t+1} = \Gamma(\boldsymbol{\omega}_t, \mathbf{z}_t, j)$ where j is the winning bidder.

It is convenient to write the maximization problem from the beginning of a period, prior to realization of the private shock and prior to realization of the contract characteristics. Then firms choose their bidding strategy so as to maximize the expected discounted sum

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t (b_{it} - c_{it}) \Big\{ b_{it} \leqslant \min_j (b_{jt}) \Big\} \Big| \boldsymbol{\omega}_0, \mathbf{z}_{-1} \right], \tag{91}$$

where $\mathbf{z}_{-1}$ refers to last period's contract and where the expectation is defined over rival's bids in all periods as well as over the firm's own costs in all periods. Due to the Markov structure, this maximization problem can be written recursively

$$\begin{aligned} V_i(\boldsymbol{\omega}_t, \mathbf{z}_{t-1}) = \iint \max_{b_{it}} \Bigg[& (b_{it} - c_{it}) \Pr(i \text{ wins} | b_{it}, \boldsymbol{\omega}_t, \mathbf{z}_t) \\ & + \beta \sum_{j=1}^{N} \Pr(j \text{ wins} | b_{it}, \boldsymbol{\omega}_t, \mathbf{z}_t) \\ & \times V_i\big(\Gamma(\boldsymbol{\omega}_t, \mathbf{z}_t, j), z_t\big) \Bigg] \mathrm{d}F(c_{it} | \boldsymbol{\omega}_t, \mathbf{z}_t)\, \mathrm{d}G(\mathbf{z}_t | \mathbf{z}_{t-1}). \end{aligned} \tag{92}$$

As is now common practice in the empirical auctions literature [Guerre, Perrigne and Vuong (2000)], JP show that bidders' costs can be recovered by inverting the first order condition associated with the optimal bid. Let $G^i(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t)$ be the distribution of bids submitted by bidder i conditional on the state variables and $g^i(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t)$ be the density function. Let

$$h^i(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t) = \frac{g^i(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t)}{1 - G^i(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t)}$$

denote the associated hazard function, and note that

$$\frac{\partial \Pr(i \text{ wins} | b_{it}, \boldsymbol{\omega}_t, \mathbf{z}_t)}{\partial b_{i,t}} = -\sum_{j \neq i} h^j(b_{it} | \boldsymbol{\omega}_t, \mathbf{z}_t) \Pr(i \text{ wins} | b_{it}, \boldsymbol{\omega}_t, \mathbf{z}_t)$$

while

$$\frac{\partial \Pr(j \text{ wins} | b_{it}, \boldsymbol{\omega}_t, \mathbf{z}_t)}{\partial b_{i,t}} = h^j(b_{it} | \boldsymbol{\omega}_t, \mathbf{z}_t) \Pr(i \text{ wins} | b_{it}, \boldsymbol{\omega}_t, \mathbf{z}_t).$$

Using these expressions, the first order condition for optimal bids yields the equation

$$\begin{aligned} b_{it} = c_{it} + \frac{1}{\sum_{j \neq i} h^j(b_{it} | \boldsymbol{\omega}_t, \mathbf{z}_t)} - \beta \sum_{j \neq i} \frac{h^j(b_{it} | \boldsymbol{\omega}_t, \mathbf{z}_t)}{\sum_{l \neq i} h^l(b_{it} | \boldsymbol{\omega}_t, \mathbf{z}_t)} \\ \times \big[V_i\big(\Gamma(\boldsymbol{\omega}_t, \mathbf{z}_t, i), \mathbf{z}_t\big) - V_i\big(\Gamma(\boldsymbol{\omega}_t, \mathbf{z}_t, j), \mathbf{z}_t\big) \big]. \end{aligned} \tag{93}$$

The optimal bid equals the cost plus a markup that has two terms. The first term reflects competition in the current auction. The second term accounts for the incremental effect on future profits of firm i winning today's auction.

Since the first order condition is strictly increasing in c it can be inverted to obtain

$$c = \phi(b|\boldsymbol{\omega}_t, \mathbf{z}_t), \tag{94}$$

where ϕ is a function of the observed bids, the hazard function of bids, h, the transition function, Γ, and the continuation values, V. The transition function is a known function. Since the bids, contract characteristics, and state variables are observed, the hazard function of bids can be obtained from the data. Thus, if the continuation values were known, then the relationship in (94) could be used to infer bidders' costs. Hence, as in the examples above, in order to estimate the parameters of the cost distribution we need first to obtain estimates of the continuation values.

3.7.1. Estimating continuation values

In order to estimate the continuation values, JP note that the continuation values can be written as a function only of the distribution of bids. The easiest way to see this is to inspect (91). The expected discounted value involves terms in the probability of winning, which can be derived from the distribution of bids, and terms in the expected markup. Equation (93) shows that the optimal markup is a function of the distribution of bids and the continuation values. JP show that by combining these two equations it is possible to write the continuation values as a function only of the distribution of bids.

The derivation is long so we omit it here and instead refer readers to the appendix of JP. Proposition 1 in JP shows that Equations (92) and (93) can be manipulated to obtain

$$\begin{aligned} V_i(\boldsymbol{\omega}_t, \mathbf{z}_{t-1}) = \int \Bigg\{ \int & \frac{1}{\sum_{j\neq i} h^j(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t)} \, \mathrm{d}G^{(i)}(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t) \\ & + \beta \sum_{j\neq i} \Bigg[\Pr(j \text{ wins}|\boldsymbol{\omega}_t, \mathbf{z}_t) \\ & + \int \frac{h^i(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t)}{\sum_{l\neq i} h^l(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t)} \, \mathrm{d}G^{(j)}(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t) \Bigg] \\ & \times V_i\big(\Gamma(\boldsymbol{\omega}_t, \mathbf{z}_t, j), \mathbf{z}_t\big) \Bigg\} \mathrm{d}G(\mathbf{z}_t|\mathbf{z}_{t-1}), \end{aligned} \tag{95}$$

where the notation

$$G^{(i)}(\cdot) = \prod_{k\neq i} \big[1 - G^k(b|\boldsymbol{\omega}_t, \mathbf{z}_t)\big] g^i(b|\boldsymbol{\omega}_t, \mathbf{z}_t).$$

The terms in square brackets in the second line of (95) sum to one and therefore can be interpreted as transition probabilities. This interpretation leads to the following

construction. Assume that the state space is discrete and let A_i be a vector with one element for each state representing the first term above

$$A_i(s) = \iint \frac{1}{\sum_{j\neq i} h^j(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t)} \mathrm{d}G^{(i)}(\cdot|\boldsymbol{\omega}_t, \mathbf{z}_t)\, \mathrm{d}G(\mathbf{z}_t|\mathbf{z}_{t-1}).$$

Next, construct the matrix M^i such that each element (k, l) reflects the transition probabilities above

$$M^i_{k,l} = \begin{cases} [\Pr(j \text{ wins}|\omega_k, z_l) + \int \frac{h^i(\cdot|\omega_k, z_l)}{\sum_{l\neq i} h^l(\cdot|\omega_k, z_l)} \mathrm{d}G^{(j)}(\cdot|\omega_k, z_l)] \Pr(z_l|z_k), \\ \quad \text{if } \omega_l = \Gamma(\omega_k, z_l, j), \\ 0, \quad \text{otherwise.} \end{cases}$$

Then the value function can be expressed as

$$V_i = \left[I - \beta M^i\right]^{-1} A_i. \tag{96}$$

The matrices M^i and A_i can be estimated using estimates of the bid distribution.

3.7.2. Estimating the cost distribution

Once the continuation values are known, estimating the cost distribution is straightforward. There is a relationship between the cost distribution and the bid distribution that is given by

$$F(c|\boldsymbol{\omega}_t, \mathbf{z}_t) = G\big(b(c, \boldsymbol{\omega}_t, \mathbf{z}_t)\big|\boldsymbol{\omega}_t, \mathbf{z}_t\big) = G\big(\phi^{-1}(c|\boldsymbol{\omega}_t, \mathbf{z}_t)\big|\boldsymbol{\omega}_t, \mathbf{z}_t\big)$$

(provided that ϕ is invertible). The function ϕ can be estimated using the first order condition (93) and the estimated continuation values. The estimated ϕ can then be substituted into the estimated bid distribution in order to obtain an estimate of the cost distribution.

3.8. Outstanding issues

The literature on structural estimation of dynamic games is relatively recent. As a result our focus has been on reviewing assumptions and techniques that make it *feasible* to use the implications of dynamic games to make inferences on parameters of interest to I.O. We have paid little attention to a host of related issues including; the asymptotic efficiency of alternative estimators, the small sample properties of those estimators, identification in the absence of auxiliary information, and the likely validity of various assumptions.

It is not our intention to minimize any of these issues. Indeed we think it important to explore all of them, particularly the assumptions underlying the analysis. This includes the behavioral assumptions and the assumptions regarding the selection of equilibria, as well as more traditional assumptions on the properties of the unobservables in the model. The simple fact is that we have little to report on most of these issues. There

is however one exception; problems that arise due to the presence of serially correlated unobserved state variables. Since this is an issue that has appeared in several related literatures, we do have some idea of how to deal with it in the context of estimating dynamic games.

3.8.1. Serially correlated unobserved state variables

In all of the examples above it is assumed that all of the states that are commonly known to the agents are also observed by the econometrician. In many empirical applications this assumption is questionable. For example, in many cases we might expect there to be an aggregate shock to profits that is known to all of the firms, but not controlled for by the econometrician. The models presented above can be modified to accommodate these shocks if they are i.i.d. over time. However we would often expect aggregate shocks to be serially correlated, just as most aggregate variables are. In that case, behavior in previous periods would depend on previous realizations of the unobserved states, leading to a correlation between today's values of the observed and unobserved states.

The statement of the problems caused by unobserved serially correlated state variables in dynamic models with discrete outcomes dates back at least to Heckman (1981). Pakes (1994) reviews three possible solutions to the problem: (i) solving for the unobserved states as a function of observables, (ii) simulating the model from a truly exogenous distribution of initial conditions, and (iii) using the ergodic distribution to model the long run relationship between the unobserved and observed states. With the advent of longer panels there is also the possibility of using techniques that allow one or more parameters to differ across markets in a panel of markets (say a market specific time invariant profit parameter, or a separate initial condition for each market), and then estimate those parameters pointwise.

The first case is quite promising in contexts where there is an observable continuous response to the unobservable state. Then conditional on the parameter vector, there is often a one to one relationship between the unobserved states and the observed states and controls. Several papers in the literature on static demand system estimation [Berry (1994), Berry, Levinsohn and Pakes (1995), and Bajari and Benkard (2005)] have used such a condition to recover serially correlated unobserved product characteristics using data on quantities, prices, and observed product characteristics. Timmins (2002) uses a similar procedure to control for the initial conditions in a single agent dynamic control problem with unobserved state variables. Olley and Pakes (1996) use the implications of a dynamic Markov perfect equilibrium model to recover a serially correlated productivity term. These methods could be used to recover the unobserved state variables prior to the dynamic estimation, and then the unobserved state variables could be treated as if they were observed in the dynamic estimation algorithm (at least up to estimation error).

Things become more difficult when the unobserved states are not recoverable in this way. In single-agent dynamic models, several papers [e.g. Pakes (1986) and Keane and

Wolpin (1997)] have used exogenous initial conditions to solve the problem of serially correlated unobserved states. Starting with an exogenous initial distribution of states, the model can be used to simulate the relationship between the observed and unobserved states in future periods. However, while there may be many reasonable ways of modelling initial conditions for a single agent (be it a firm or an individual), such conditions tend to be less realistic for an industry, whose history is typically much longer than the available data.

The third case is perhaps more appropriate for industry studies. Instead of using an exogenous initial condition for the unobserved states at the time the market starts up, we assume that the draws on the joint distribution of the unobserved states and the initial condition are draws from an invariant distribution. That distribution is then estimated along with the other parameters of the problem. The rational here is that if the markets in question have been in existence long enough, the joint distribution of the initial condition and the unobserved state will not depend on the early years of the industry's evolution. Rather it will depend only on the limiting structure of the Markov process generated by the nature of the dynamic game, a structure we can analyze.

Aguirregabiria and Mira (2007) implement one version of this solution. They allow for an unobserved fixed effect which varies across markets and assume both that the fixed effect can only take on a finite number of values and that the transition probabilities for the observed exogenous variables are independent of the values of the fixed effect. They then solve for an invariant distribution of the state of the system and the fixed effect, form the conditional distribution of the initial condition given the fixed effect, and integrate out over possible values of the fixed effect. They report that allowing for the fixed effect has a noticeable impact on their empirical results.

Of course, if one has a reasonably long panel of markets one should be able to estimate the fixed effect (or some other unobserved initial condition) pointwise; our fourth solution possibility. In that case continuation values could be estimated as described above but separately for each market in the data. The observations across markets could then be pooled together in the second stage in order to estimate the structural parameters that are the same across markets. This would lead to substantially higher estimation error in the continuation values, and one might want to think hard about estimators that would be designed to minimize the impact of these errors.[62] Some Monte Carlo work on just how long a panel is likely to be required for this procedure to be fruitful would be extremely helpful.

References

Abreu, D., Pearce, D., Stacchetti, E. (1986). "Optimal cartel equilibria with imperfect monitoring". Journal of Economic Theory 39 (1), 251–259.

[62] Note that this process could also solve the multiple equilibrium problem that might exist across markets.

Ackerberg, D. (2003). "Advertising learning and customer choice in experience good markets: A structural empirical examination". International Economic Review 44 (3), 1007–1040.
Ackerberg, D., Pakes, A. (2005). "Notes on relaxing the scalar unobservable assumption in Olley and Pakes". Mimeo. Harvard University.
Ackerberg, D., Caves, K., Fraser, G. (2004). "Structural identification of production functions". Mimeo. UCLA.
Aguirregabiria, V., Mira, P. (2007). "Sequential estimation of dynamic games". Econometrica 75 (1), 1–54.
Ahn, H., Powell, J. (1993). "Semiparametric estimation of censored selection models with a nonparametric selection mechanism". Journal of Econometrics 58 (1–2), 3–29.
Arellano, M., Bond, S. (1991). "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations". Review of Economic Studies 58, 277–297.
Arellano, M., Bover, O. (1995). "Another look at the instrumental variable estimation of error-components models". Journal of Econometrics 68 (1), 29–51.
Bajari, P., Benkard, C.L. (2005). "Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach". Journal of Political Economy 113 (6), 1239–1276.
Bajari, P., Benkard, C.L., Levin, J. (2007). "Estimating dynamic models of imperfect competition". Econometrica. In press.
Benkard, C.L. (2000). "Learning and forgetting: The dynamics of aircraft production". American Economic Review 90 (4), 1034–1054.
Benkard, C.L. (2004). "A dynamic analysis of the market for wide-bodied commercial aircraft". Review of Economic Studies 71 (3), 581–612.
Berry, S. (1994). "Estimating discrete choice models of product differentiation". RAND Journal of Economics 25 (2), 242–262.
Berry, S., Pakes, A. (2002). "Estimation from the optimality conditions for dynamic controls". Mimeo. Yale University.
Berry, S., Pakes, A. (2005). "The pure characteristics demand model". Mimeo. Yale University.
Berry, S., Levinsohn, J., Pakes, A. (1995). "Automobile prices in market equilibrium". Econometrica 63 (4), 841–890.
Berry, S., Levinsohn, J., Pakes, A. (1999). "Voluntary export restraints on automobiles: Evaluating a trade policy". American Economic Review 89 (3), 400–430.
Berry, S., Levinsohn, J., Pakes, A. (2004). "Estimating differentiated product demand systems from a combination of micro and macro data: The new car model". Journal of Political Economy 112 (1), 68–105.
Berry, S., Linton, O., Pakes, A. (2004). "Limit theorems for estimating the parameters of differentiated product demand systems". Review of Economics Studies 71 (3), 613–654.
Blundell, R., Bond, S. (1999). "GMM estimation with persistent panel data: An application to production functions". Working Paper Series No. W99/4. The Institute for Fiscal Studies.
Bond, S., Söderbom, M. (2005). "Adjustment costs and the identification of Cobb Douglas production functions". Institute for Fiscal Studies WP O5/04.
Bresnahan, T. (1981). "Departures from marginal-cost pricing in the American automobile industry: Estimates for 1977–1978". Journal of Econometrics 17 (2), 201–227.
Bresnahan, T. (1982). "The oligopoly solution concept is identified". Economics Letters 10, 87–92.
Bresnahan, T. (1987). "Competition and collusion in the American automobile industry: The 1955 price war". Journal of Industrial Economics 35 (4), 457–482.
Buettner, T. (2004a). "Productivity dynamics when labor is a state variable". Mimeo. LSE.
Buettner, T. (2004b). "R&D and the dynamics of productivity". Mimeo. LSE.
Caballero, R., Engel, E.M.R.A. (1999). "Explaining investment dynamics in US manufacturing: A generalized (S, s) approach". Econometrica 67 (4), 783–826.
Chamberlain, G. (1984). "Panel data". In: Arrow, K., Intriligator, M. (Eds.), Handbook of Econometrics, vol. 2. Elsevier Science, pp. 1247–1318. Chapter 22.
Chan, T. (2002). "Demand for soft drinks: Characteristics, corners and continuous choice". Mimeo. Washington University, Marketing Department.

Chernozhukov, V., Hong, H., Tamer, E. (2007). "Estimation and confidence regions for parameter sets in economic models". Econometrica. In press.
Court, A. (1939). "Hedonic price indexes with automotive examples". In: The Dynamics of Automobile Demand. General Motors Corporation, Detroit, pp. 99–117.
Crawford, G., Shum, M. (2007). "Uncertainty and learning in pharmaceutical demand". Econometrica 73 (4), 1137–1173.
Das, M., Olley, G.S., Pakes, A. (1996). "The evolution of the market for consumer electronics". Mimeo. Harvard University.
Davis, S., Haltiwanger, J. (1992). "Gross job creation, gross job destruction, and employment reallocation". Quarterly Journal of Economics 107 (2), 819–864.
Deaton, A., Muellbauer, J. (1980). "An almost ideal demand system". American Economic Review 70 (3), 312–326.
Department of Justice. (1992). "US Merger guidelines".
Doraszelski, U., Judd, K. (2004). "Solution methods for Markov perfect Nash equilibria of continuous time, finite state stochastic games". Mimeo. Hoover Institution.
Doraszelski, U., Markovich, S. (2007). "Advertising dynamics and competitive advantage". RAND Journal of Economics. In press.
Doraszelski, U., Pakes, A. (2007). "A framework for applied dynamic analysis in I.O." In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization, vol. 3. North-Holland, Amsterdam. In press.
Doraszelski, U., Satterthwaite, M. (2007). "Foundations of Markov-perfect industry dynamics: Existence, purification, and multiplicity". Mimeo. Hoover Institution.
Dubin, J., McFadden, D. (1984). "An econometric analysis of residential electric appliance holdings and consumption". Econometrica 52 (2), 345–362.
Dunne, T., Roberts, M., Samuelson, L. (1988). "Patterns of firm entry and exit in US manufacturing industries". RAND Journal of Economics 19 (4), 495–515.
Dunne, T., Klimek, S., Roberts, M., Xu, Y. (2006). "Entry and exit in geographic markets". Mimeo. Pennsylvania State University.
Einav, L. (2003). "Not all rivals look alike: Estimating an equilibrium model of the release date timing game". Mimeo. Stanford University.
Erdem, T., Keane, M. (1996). "Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets". Marketing Science 15 (1), 1–20.
Erdem, T., Imai, S., Keane, M. (2003). "Brand and quantity choice dynamics under price uncertainty". Quantitative Marketing and Economics 1 (1), 5–64.
Ericson, R., Pakes, A. (1995). "Markov-perfect industry dynamics: A framework for empirical work". Review of Economic Studies 62 (1), 53–83.
Esteban, S., Shum, M. (2002). "Durable good monopoly with secondary markets: The case of automobiles". Mimeo. Johns Hopkins.
Fershtman, C., Pakes, A. (2000). "A dynamic oligopoly with collusion and price wars". RAND Journal of Economics 31 (2), 207–236.
Fershtman, C., Pakes, A. (2005). "Finite state dynamic games with asymmetric information: A computational framework". Mimeo. Harvard University.
Gentzkow, M. (2004). "Valuing new goods in a model with complementarity: Online newspapers". Mimeo. Chicago GSB.
Gorman, W. (1959). "Separable utility and aggregation". Econometrica 27, 469–481.
Gowrisankaran, G. (1995). "A dynamic analysis of mergers". PhD dissertation, Department of Economics, Yale University.
Gowrisankaran, G., Town, R. (1997). "Dynamic equilibrium in the hospital industry". Journal of Economics and Management Strategy 6 (1), 45–74.
Green, E., Porter, R. (1984). "Non-cooperative collusion under imperfect price information". Econometrica 52, 87–100.
Greenstreet, D. (2005). "Exploiting sequential learning to estimate establishment-level productivity dynamics and decision rules". Mimeo. University of Michigan.

Griliches, Z. (1961). "Hedonic price indexes for automobiles: An econometric analysis of quality change". In: The Price Statistics of the Federal Government. NBER, New York.
Griliches, Z., Hausman, J. (1986). "Errors in variables in panel data". Journal of Econometrics 31 (1), 93–118.
Gronau, R. (1974). "Wage comparisons – A selectivity bias". Journal of Political Economy 82 (6), 1119–1143.
Guerre, E., Perrigne, I., Vuong, Q. (2000). "Optimal nonparametric estimation of first-price auctions". Econometrica 68 (3), 525–574.
Hansen, L. (1982). "Large sample properties of generalized method of moments estimators". Econometrica 50 (4), 1029–1054.
Hausman, J. (1996). "Valuation of new goods under perfect and imperfect competition". In: Bresnahan, T., Gordon, R. (Eds.), The Economics of New Goods. University of Chicago Press, Chicago, pp. 209–248. Chapter 5.
Heckman, J. (1974). "Shadow prices, market wages, and labor supply". Econometrica 42 (4), 679–694.
Heckman, J. (1976). "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models". Annals of Economic and Social Measurement 5 (4), 475–492.
Heckman, J. (1978). "Dummy endogenous variables in a simultaneous equation system". Econometrica 46 (4), 931–959.
Heckman, J. (1979). "Sample selection bias as a specification error". Econometrica 47 (1), 153–161.
Heckman, J. (1981). "The incidental parameters problem and the problem of initial conditions in estimating a discrete time discrete data stochastic process". In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data and Econometric Applications. MIT press. Chapter 4.
Heckman, J., Honoré, B. (1990). "The empirical content of the roy model". Econometrica 58, 1121–1149.
Heckman, J., Robb, R. (1985). "Alternative methods for evaluating the impact of interventions: An overview". Journal of Econometrics 30 (1–2), 239–267.
Heckman, J., Snyder, J.M. (1997). "Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators". RAND Journal of Economics 28 (0), S142–S189.
Hendel, I. (1999). "Estimating multiple-discrete choice models: An application to computerization returns". Review of Economic Studies 66 (2), 423–446.
Hendel, I., Nevo, A. (2002). "Sales and consumer inventory". NBER working paper #9048.
Hoch, I. (1962). "Estimation of production parameters combining time-series and cross-section data". Econometrica 30 (1), 34–53.
Houthakker, H.S. (1955). "The Pareto distribution and the Cobb–Douglas production function". Review of Economic Studies 23 (1), 27–31.
Hotelling, H. (1929). "Stability in competition". Economic Journal 39, 41–57.
Hotz, V., Miller, R. (1993). "Conditional choice probabilities and the estimation of dynamic models". Review of Economic Studies 60 (3), 497–529.
Hotz, V., Miller, R., Sanders, S., Smith, J. (1994). "A simulation estimator for dynamic models of discrete choice". Review of Economic Studies 60, 397–429.
Imbens, G., Lancaster, T. (1994). "Combining micro and macro data in microeconometric models". Review of Economic Studies 61 (4), 655–680.
Jofre-Bonet, M., Pesendorfer, M. (2003). "Estimation of a dynamic auction game". Econometrica 71 (5), 1443–1489.
Keane, M., Wolpin, K. (1997). "The career decisions of young men". Journal of Political Economy 105 (3), 473–522.
Lancaster, K. (1971). Consumer Demand: A New Approach. Columbia University Press, New York.
Lau, L.J. (1982). "On identifying the degree of competitiveness from industry price and output data". Economics Letters 10, 93–99.
Levinsohn, J., Petrin, A. (2003). "Estimating production functions using inputs to control for unobservables". Review of Economic Studies 70 (2), 317–341.
Magnac, T., Thesmar, D. (2002). "Identifying dynamic discrete decision processes". Econometrica 70 (2), 801–816.

Manski, C., Lerman, S. (1977). "The estimation of choice probabilities from choice based samples". Econometrica 45 (8), 1977–1988.
Marschak, J., Andrews, W.H. (1944). "Random simultaneous equations and the theory of production". Econometrica 12 (3–4), 143–205.
McFadden, D. (1974). "Conditional logit analysis of qualitative choice behavior". In: Zarembka, P. (Ed.), Frontiers of Econometrics. Academic Press, New York.
McFadden, D. (1981). "Econometric models of probabilistic choice". In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA.
Mundlak, Y. (1961). "Empirical production function free of management bias". Journal of Farm Economics 43 (1), 44–56.
Nadiri, M.I., Rosen, S. (1974). A Disequilibrium Model of Demand for Factors of Production. National Bureau of Economic Research, New York.
Nevo, A. (2001). "Measuring market power in the ready-to-eat cereal industry". Econometrica 69 (2), 307–342.
Olley, G.S., Pakes, A. (1996). "The dynamics of productivity in the telecommunications equipment industry". Econometrica 64 (6), 1263–1298.
Pakes, A. (1986). "Patents as options: Some estimates of the value of holding European patent stocks". Econometrica 54, 755–784.
Pakes, A. (1994). "The estimation of dynamic structural models: Problems and prospects, part II. Mixed continuous-discrete control models and market interactions". In: Laffont, J.J., Sims, C. (Eds.), Advances in Econometrics: Proceedings of the 6th World Congress of the Econometric Society. Cambridge Univ. Press. Chapter 5.
Pakes, A. (1995). "Biases in the CPI". Testimony before the Senate Finance Committee, the United States Congress.
Pakes, A. (1998). "A framework for dynamic analysis in applied I.O.". NBER discussion paper No. 8024.
Pakes, A. (2004). "Common sense and simplicity in industrial organization". Review of Industrial Organization 23, 193–215.
Pakes, A., Griliches, Z. (1984). "Estimating distributed lags in short panels with an application to the specification of depreciation patterns and capital stock constructs". Review of Economic Studies 51, 243–262.
Pakes, A., McGuire, P. (1994). "Computing Markov-perfect Nash equilibria: Numerical implications of a dynamic differentiated product model". RAND Journal of Economics 25 (4), 555–589.
Pakes, A., McGuire, P. (2001). "Stochastic approximation for dynamic models: Markov perfect equilibrium and the 'Curse' of dimensionality". Econometrica 69 (5), 1261–1281.
Pakes, A., Olley, G.S. (1995). "A limit theorem for a smooth class of semiparametric estimators". Journal of Econometrics 65 (1), 295–332.
Pakes, A., Pollard, D. (1989). "Simulation and the asymptotics of optimization estimators". Econometrica 57 (5), 1027–1057.
Pakes, A., Ostrovsky, M., Berry, S. (2007). "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)". RAND Journal of Economics. In press.
Pakes, A., Porter, J., Ho, K., Ishii, J. (2006). "Moment inequalities and their application". Mimeo. Harvard University.
Pavcnik, N. (2002). "Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants". Review of Economic Studies 69, 245–276.
Pesendorfer, M., Schmidt-Dengler, P. (2003). "Identification and estimation of dynamic games". NBER working paper #9726.
Petrin, A. (2002). "Quantifying the benefits of new products: The case of the minivan". Journal of Political Economy 110 (4), 705–729.
Robinson, P. (1988). "Root-N-consistent semiparametric regression". Econometrica 56 (4), 931–954.
Rosenbaum, P., Rubin, D. (1983). "The central role of the propensity score in observational studies for causal effects". Biometrika 70 (1), 41–55.
Rotemberg, J., Saloner, G. (1985). "A supergame-theoretic model of business cycles and price wars during booms". American Economic Review 76, 390–407.

Rubinstein, R. (1981). Simulation and the Monte Carlo Method. John Wiley and Sons, New York.
Rust, J. (1987). "Optimal replacement of GMC bus engines". Econometrica 55, 999–1033.
Rust, J. (1994). "Structural estimation of Markov decision processes". In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. IV. Elsevier Science, Amsterdam.
Ryan, S. (2006). "The costs of environmental regulation in a concentrated industry". Mimeo. Duke University.
Salop, S. (1979). "Monopolistic competition with outside goods". RAND Journal of Economics 10 (1), 141–156.
Shaked, A., Sutton, J. (1982). "Relaxing price competition through product differentiation". Review of Economic Studies 49 (1), 3–13.
Schmidt-Dengler, P. (2003). "The timing of new technology adoption: The case of MRI". Mimeo. Yale University.
Sobel, J. (1984). "The timing of sales". Review of Economic Studies 51 (3), 353–368.
Song, M. (2004). "Measuring consumer welfare in the CPU market: An application of the pure characteristics demand model". Mimeo. Georgia Tech.
Timmins, C. (2002). "Measuring the dynamic efficiency costs of regulators' preferences: Municipal water utilities in the arid west". Econometrica 70 (2), 603–629.
Train, K. (2003). Discrete Choice Models with Simulation. Cambridge University Press.
Weintraub, G., Benkard, C.L., Van Roy, B. (2007). "Markov perfect industry dynamics with many firms". Mimeo. Stanford University.
Wooldridge, J. (2002). Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.
Wooldridge, J. (2004). "On estimating firm-level production functions using proxy variables to control for unobservables". Mimeo. Michigan State University.

Chapter 64

STRUCTURAL ECONOMETRIC MODELING: RATIONALES AND EXAMPLES FROM INDUSTRIAL ORGANIZATION

PETER C. REISS

Graduate School of Business, Stanford University, Stanford, CA 94305-5015, USA
e-mail: preiss@optimum.stanford.edu

FRANK A. WOLAK

Department of Economics, Stanford University, Stanford, CA 94305-6072, USA
e-mail: wolak@zia.stanford.edu

Contents

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06064-3

Abstract

This chapter explains the logic of structural econometric models and compares them to other types of econometric models. We provide a framework researchers can use to develop and evaluate structural econometric models. This framework pays particular attention to describing different sources of unobservables in structural models. We use our framework to evaluate several literatures in industrial organization economics, including the literatures dealing with market power, product differentiation, auctions, regulation and entry.

Keywords

structural econometric model, market power, auctions, regulation, entry

JEL classification: C50, C51, C52, D10, D20, D40

1. Introduction

The founding members of the Cowles Commission defined *econometrics* as: "a branch of economics in which economic theory and statistical method are fused in the analysis of numerical and institutional data" [Hood and Koopmans (1953, p. xv)]. Today economists refer to models that combine explicit economic theories with statistical models as *structural econometric models*.

This chapter has three main goals. The first is to explain the logic of structural econometric modeling. While structural econometric models have the logical advantage of detailing the economic and statistical assumptions required to estimate economic quantities, the fact that they impose structure does not automatically make them sensible. To be convincing, structural models minimally must be: (1) flexible statistical descriptions of data; (2) respectful of the economic institutions under consideration; and, (3) sensitive to the nonexperimental nature of economic data. When, for example, there is little economic theory on which to build, the empiricist may instead prefer to use nonstructural or descriptive econometric models. Alternatively, if there is a large body of relevant economic theory, then there may significant benefits to estimating a structural econometric model – provided the model can satisfy the above demands.

A second goal of this chapter is to describe the ingredients of structural models and how structural modelers go about evaluating them. Our discussion emphasizes that the process of building a structural model involves a series of related steps. These steps are by no means formulaic and often involve economic, statistical and practical compromises. Understanding when and why structural modelers must make compromises, and that structural modelers can disagree on compromises, is important for understanding that structural modeling is in part "art". For example, structural modelers often introduce "conditioning variables" that are not explicitly part of the economic theory as a way of controlling for plausible differences across observations.

Our third goal is to illustrate how structural modeling tradeoffs are made in practice. Specifically, we examine different types of structural econometric models developed by industrial organization ("IO") economists. These models examine such issues as: the extent of market power possessed by firms; the efficiency of alternative market allocation mechanisms (e.g., different rules for running single and multi-unit auctions); and the empirical implications of information and game-theoretic models. We should emphasize that this chapter is NOT a comprehensive survey of the IO literature or even a complete discussion of any single topic. Readers interested in a comprehensive review of a particular literature should instead begin with the surveys we cite. Our goal is instead to illustrate selectively how IO researchers have used economic and statistical assumptions to identify and estimate economic magnitudes. Our hope is that in doing so, we can provide a better sense of the benefits and limitations of structural econometric models.

We begin by defining structural econometric models and discussing when one would want to use a structural model. As part of this discussion, we provide a framework

for evaluating the benefits and limitations of structural models. The remainder of the chapter illustrates some of the practical tradeoffs IO researchers have made.

2. Structural models defined

In structural econometric models, economic theory is used to develop mathematical statements about how a set of observable "endogenous" variables, y, are related to another set of observable "explanatory" variables, x. Economic theory also may relate the y variables to a set of unobservable variables, ξ. These theoretical relations usually are in the form of equalities: $y = g(x, \xi, \Theta)$, where $g(\cdot)$ is a known function and Θ a set of unknown parameters or functions. Occasionally, economic theory may only deliver inequality relations, such as $y \geqslant g(x, \xi, \Theta)$.

Economic theory alone usually does not provide enough information for the econometrician to estimate Θ. For this reason, and because the economic model $y = g(x, \xi, \Theta)$ may not be able to rationalize the observed data perfectly, the econometrician adds statistical assumptions about the joint distribution of x, ξ and other unobservables appended to the model. Taken together, these economic and statistical assumptions define an empirical model that is capable of rationalizing all possible observable outcomes. In order to estimate the underlying primitives of this model, researchers use statistical objects based on the model, such as a log-likelihood function for the data, $\ell(y, x \mid \Theta)$, or conditional moments, such as $E(y \mid x, \Theta)$.

Nonstructural empirical work in economics may or may not be based on formal statistical models. At one end of the spectrum are measurement studies that focus on constructing and summarizing data, such as labor force participation and unemployment rates. At the other end are those that use formal statistical models, such as autoregressive conditional volatility models. Both types of nonstructural empirical work have a long and respected tradition in economics. An excellent early example is Engel's (1857) work relating commodity budget shares to total income. Engel's finding that expenditure shares for food were negatively related to the logarithm of total household expenditures has shaped subsequent theoretical and empirical work on household consumption behavior [see Deaton and Muellbauer (1980) and Pollak and Wales (1992)]. A somewhat more recent example of descriptive work is the Phillips curve. Phillips (1958) documented an inverse relationship between United Kingdom unemployment rates and changes in wage rates. This work inspired others to document relationships between unemployment rates and changes in prices. In the ensuing years, many economic theories have been advanced to explain why Phillips curves are or are not stable economic relations.

Nonstructural empirical models usually are grounded in economics to the extent that economics helps identify which variables belong in y and which in x. This approach, however, ultimately estimates characteristics of the joint population density of x and y, $f(x, y)$, or objects that can be derived from it, such as:

$f(y \mid x)$, the conditional density of y given x;
$E(y \mid x)$, the conditional expectation of y given x;
$\mathrm{Cov}(y \mid x)$, the conditional covariances (or correlations) of y given x; or,
$Q_\alpha(y \mid x)$ the α conditional quantile of y given x.

The most commonly estimated characteristic of the joint density is the best linear predictor (BLP($y \mid x$)) of y given x.

More recently researchers have taken advantage of developments in nonparametric and semiparametric statistical methods to derive consistent estimates of the joint density of y and x. For example, statisticians have proposed kernel density techniques and other data smoothing methods for estimating $f(x, y)$. These same smoothing techniques have been used to develop nonparametric conditional mean models. Silverman (1986), Härdle (1990), Härdle and Linton (1994) and others provide useful introductions to these procedures. A major advantage of nonparametric models is that they can provide flexible descriptions of the above statistical quantities.

Given their flexibility, it would seem that nonstructural empirical researchers should always prefer nonparametric methods. There are, however, limitations to nonparametric methods. One is that they may require large amounts of data to yield much precision.[1] Second, and more important, once estimated, it is unclear how a flexible estimate of a joint density can be used to recover economic constructs such as economies of scale in production and consumer welfare. Moreover, it is also unclear how to perform out-of-sample counterfactual calculations, such as the impact of an additional bidder on the winning bid in an auction.

It is tempting to look at our descriptions of structural versus nonstructural models, and parametric versus nonparametric models, and see them as absolutes – empirical models are either structural or nonstructural, parametric or nonparametric. We see little value in such absolute classification exercises. In practice, it is not uncommon to find structural econometric models that include nonstructural components or nonparametric components. Our goal in providing these definitions is to have an initial basis for classifying and evaluating the success of an econometric model.

An initial example from IO may help understand our focus and intent. Consider a researcher who observes the winning bids, $y = \{y_1, \ldots, y_T\}'$, in each of a large number of T similar auctions. Suppose the researcher also observes the number of bidders, $x = \{x_1, \ldots, x_T\}'$, in each auction. To understand the equilibrium relationship between winning bids and the number of bidders the researcher could use a structural or a nonstructural approach.

[1] Silverman (1986) argues that researchers using these techniques face a "curse of dimensionality", wherein the amount of data required to obtain precise estimates grows rapidly with the dimensions of x and y. His calculations (1986, Table 4.2) suggest that researchers may need thousands, if not hundreds of thousands of observations before they can place great faith in these flexible techniques. For instance, more than ten times as much data is required to attain the same level of precision for a four-dimensional as a two-dimensional joint density. More than 200 times as much data is required for an eight-dimensional as a four-dimensional density.

A standard nonstructural approach would be to regress winning bids on the number of bidders. Under standard statistical assumptions, this regression would deliver the best linear predictor of winning bids given the number of bidders. These coefficient estimates could be used to predict future winning bids as a function of the number of bidders. Alternatively, a researcher worried about a nonlinear relationship between winning bids and the number of bidders might instead opt to use nonparametric smoothing techniques to estimate the conditional density of winning bids given each distinct observed number of bidders x, $f(y \mid x)$. The researcher could then use this estimated conditional density, $\widehat{f(y \mid x)}$, to calculate whether, for example, expected winning bids in the sample auctions increased or decreased with the number of bidders. The researcher could also check to see if the conditional expected bid increased or decreased linearly with the number of bidders.

The process of formulating and implementing either of these nonstructural models so far has made little use of economics, except perhaps to identify what is y and what is x. For instance, our discussion of these descriptive models has made no reference to institutional features of the auctions (e.g., sealed-bid versus open-outcry and first-price versus second-price). It also has not required economic assumptions about bidder behavior or characteristics (e.g., risk-neutrality, expected profit maximization and bidder competition). In some cases (e.g., sealed-bid versus open-outcry), we may be able to incorporate these considerations into a nonstructural model by introducing them as conditioning variables. In other cases (e.g., the degree of risk aversion), this may not be possible.

A key reason then to use economic theory, beyond specifying x and y, is to clarify how institutional and economic conditions affect relationships between y and x. This specificity is essential once the researcher wishes to make causal statements about estimated relationships or use them to perform counterfactuals. Suppose for example, the researcher has regressed winning bids on the number of bidders and estimates the coefficient on the number of bidders is \$100. Can this estimate be interpreted as the causal effect of adding another bidder to a future auction? We would argue that without further knowledge about the institutional and economic features of the auctions under study the answer is no. What separates structural models from nonstructural models, and some structural models from others, is how clearly the connections are made between institutional, economic, and statistical assumptions and the estimated relationships. While it is possible to assert that assumptions exist that make the estimated relationship causal, the plausibility of such claims ultimately rests on whether these assumptions are reasonable for the researcher's application.

As we discuss later in Section 8, IO economists have developed a variety of structural models of auction bid data. These models have been used to derive causal models of the equilibrium relations between winning bids and the number of bidders. Paarsch (1992), for example, was the first to compare empirical models of winning bids in private value and common value sealed-bid auctions. For instance, he showed that for sealed-bid auctions with risk-neutral, expected profit-maximizing, Pareto-distributed-private-value bidders would have the following density of winning bids y given a known number of

bidders, x:

$$f(y \mid x, \theta) = \frac{\theta_2 x}{y^{\theta_2 x+1}} \left[\frac{\theta_1 \theta_2 (x-1)}{\theta_2 (x-1) - 1} \right]^{\theta_2 x}.$$

Using this density, Paarsch derives the expected value of the winning bid conditional on the number of bidders:

$$E(y \mid x) = \left[\frac{\theta_1 \theta_2 (x-1)}{\theta_2 (x-1) - 1} \right] \frac{\theta_2 x}{\theta_2 x - 1}.$$

Paarsch's paper motivated IO economists to think about what types of institutional, economic and statistical assumptions were necessary to recover causal relationships from auction data. For example, researchers have asked how risk aversion, collusion and asymmetric information change the equilibrium distribution of bids. Researchers also have compared the observable implications of using sealed-bid versus open-outcry auctions.

In closing this section, we should re-emphasize the general goal of structural econometric modeling. Structural econometric models use economic and statistical assumptions to identify economic quantities from the joint density of economic data, $f(x, y)$. The main strength of this approach is that, done right, it can make clear what economic assumptions are required to draw causal economic inferences from the distribution of economic data.

3. Constructing structural models

Having introduced the concept of a structural model, we now explore how structural modelers go about constructing econometric models.

3.1. Sources of structure

There are two general sources of "structure" in structural models – economics and statistics. Economics allows the researcher to infer how economic behavior and institutions affect relationships between a set of economic conditions x and outcomes y. Often these economic models are deterministic, and as such do not speak directly to the distribution of noisy economic data. Structural econometric modelers thus must add statistical structure in order to rationalize why economic theory does not perfectly explain data. As we show later, this second source of structure may affect which economic quantities a researcher can recover and which estimation methods are preferable.

In any structural modeling effort a critical issue will be: How did the structural modeler know what choices to make when introducing economic and statistical assumptions? Most answers to this question fall into one of three categories: those made to reflect economic realities; those made to rationalize what is observed in the data or describe how the data were generated; and, those made to simplify estimation. We should

note at the outset that there is no necessary agreement among structural modelers as to how to make these choices. Some purists, for example, believe that structural models must come from fully-specified stochastic economic models. Others find it acceptable to add structure if that structure facilitates estimation or allows the researcher to recover economically meaningful parameters. For instance, economic theory may make predictions about the conditional density of y given x, $f(y \mid x)$, but may be silent about the marginal density of x, $f(x)$. In this case, a researcher might assume that the marginal density of x does not contain parameters that appear in the conditional density. Of course, there is nothing to guarantee that assumptions made to facilitate estimation are reasonable.

The "structure" in a structural model is there because the researcher explicitly or implicitly chose to put it there. Although we have argued that one of the advantages of a structural econometric model is that researchers can examine the sensitivity of the structural model estimates to alternative assumptions, this is sometimes easier said than done.

The following example illustrates how some of these issues can arise even in a familiar linear regression setting. Specifically, we ask what types of assumptions are required to interpret a regression of outputs on inputs as a production function.

EXAMPLE 1. Imagine an economist with cross-section, firm-level data on output, Q_t, labor inputs, L_t, and capital inputs, K_t. To describe the relationship between firm i's output and inputs, the economist might estimate the regression:

$$\ln Q_t = \theta_0 + \theta_1 \ln L_t + \theta_2 \ln K_t + \epsilon_t, \tag{1}$$

by ordinary least squares (OLS). In this regression, the θ's are unknown coefficients and the ϵ_t is an error term that accounts for the fact that the right-hand side input variables do not perfectly predict log output.

What do we learn by estimating this regression? Absent more information we have estimated a descriptive regression. More precisely, we have estimated the parameters of the best linear predictor of $y_t = \ln(Q_t)$ given $x_t = (1, \ln(L_t), \ln(K_t))'$ for our sample of data. Goldberger (1991, Chapter 5) provides an excellent discussion of best linear predictors. The best linear predictor of y given a univariate x is $\text{BLP}(y \mid x) = a + bx$, where $a = E(y) - bE(x)$ and $b = \text{Cov}(y, x)/\text{Var}(x)$. Absent more structure, the coefficients a and b are simply functions of population moments of $f(x, y)$.

If we add to our descriptive model the assumption that the sample second moments converge to their population counterparts

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} x_t x_t' = M_{xx} \quad \text{and} \quad \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} x_t y_t = M_{xy},$$

and that M_{xx} is a matrix of full rank, then OLS will deliver consistent estimates of the parameters of the best linear predictor function. Thus, if we are interested in predicting the logarithm of output, we do not need to impose any economic structure and very little

statistical structure to estimate consistently the linear function of the logarithm of labor and logarithm of capital that best predicts (in a minimum-mean-squared-error sense) the logarithm of output.

Many economists, however, see regression (1) as more than a descriptive regression. They base their reasoning on the observation that (1) essentially looks like a logarithmic restatement of a Cobb–Douglas production function: $Q_t = A\, L_t^{\alpha} K_t^{\beta} \exp(\epsilon_t)$. Because of the close resemblance, they might interpret (1) as a production function.

A critical missing step in this logic is that a Cobb–Douglas production function typically is a deterministic relationship for the producer, whereas the regression model (1) includes an error term. Where did the error term in the empirical model come from? The answer to this question is critical because it affects whether OLS will deliver consistent estimates of the parameters of the Cobb–Douglas production function, as opposed to consistent estimates of the parameters of the best linear predictor of the logarithm of output given the logarithms of the two inputs. In other words, it is the combination of an economic assumption (production is truly Cobb–Douglas) and statistical assumptions (ϵ_t satisfies certain moment conditions) that distinguishes a structural model from a descriptive one.

Deterministic production function models provide no guidance about the properties of the disturbance in (1). The researcher thus is left to sort out what properties are appropriate from the details of the application. One could imagine, for instance, the modeler declaring that the error is an independently-distributed, mean-zero measurement error in output, and that these errors are distributed independently of the firms' input choices. In this case, OLS has the potential to deliver consistent estimates of the production function parameters.

But how did the modeler know that ϵ_t was all measurement error? As we discuss later this is likely too strong an assumption. A more plausible assumption is that the error also includes an unobservable (to the econometrician) difference in each firm's productivity (e.g., an unobservable component of A_t in the Cobb–Douglas function). The existence of such components raises the possibility that the input choices are correlated with ϵ_t. Such correlations invalidate the use of OLS to recover consistent estimates of the parameters of the Cobb–Douglas production function.

Even if one were willing to assume that ϵ_t is measurement error distributed independently of x_t, additional economic structure is necessary to interpret the OLS parameter estimates as coefficients of a Cobb–Douglas production function. By definition, a production function gives the maximum technologically feasible amount of output that can be produced from a vector of inputs. Consequently, under this stochastic structure, unless the researcher is also willing to assert that the firms in the sample are producing along their Cobb–Douglas production function, OLS applied to (1) does not yield consistent estimates of the parameters of this production function. In a theoretical realm, the assumption of technologically efficient production is relatively innocuous. However, it may not fit the institutional realities of many markets. For example, a state-owned firm may use labor in a technologically inefficient manner to maximize its political capital with unions. Regulators also may force firms to operate off their production functions.

Both of these examples underscore the point that care must exercised to ensure that the economic model fits the institutional reality of what is being studied.

This example demonstrates what assumptions are necessary for a linear regression model to have a causal economic interpretation as a production function. First, the researcher must specify an economic model of the phenomenon under consideration, including in this case the functional form for the production function. Second, she must incorporate unobservables into the economic model. This second step should receive significantly more attention than it typically does. This is because the assumptions made about the unobservables will impact the consistency of OLS parameter estimates. Sections 4 and 5 further illustrate the importance of stochastic specifications and potential pitfalls.

3.2. Why add structure?

We see three general reasons for specifying and estimating a structural econometric model.

First, a structural model can be used to estimate unobserved economic or behavioral parameters that could not otherwise be inferred from nonexperimental data. Examples of structural parameters include: marginal cost; returns to scale; the price elasticity of demand; and the impact of a change in an exogenous variable on the amount demanded or on the amount supplied.

Second, structural models can be used to perform counterfactuals or policy simulations. In counterfactuals, the researcher uses the estimated structural model to predict what would happen if elements of the economic environment change. For example, suppose that we have estimated the demand for a product and the monopolist's cost function. We could then, with added assumptions, use these estimates to calculate how market prices and quantities would change if an identical second firm entered the monopoly market.

For these calculations to be convincing, the structural modeler must be able to argue that the structural model will be invariant to the contemplated change in economic environment. Thus, if we were to evaluate instead the effect of a regulator capping a monopolist's price, we would have to maintain that the monopolist's cost function would not change as a result of the regulation. That these assumptions need to be made and checked is again another illustration of the value of a structural model – it can help researchers identify what assumptions are required in order to draw inferences and make predictions about economic behavior and phenomena.

Finally, structural models can be used to compare the predictive performance of two competing theories. For example, we could compare the performance of quantity-setting versus price-setting models of competition. It is important to emphasize that these comparisons do not provide unambiguous tests of the underlying economic theories. Indeed, these comparisons are always predicated on untestable assumptions that are not part of the theory. For instance, any "test" of quantity-setting behavior versus price-setting

behavior is predicated on the maintained functional forms for demand, costs, and the unobservables. Thus, the only sense in which one can "test" the two theories is to ask whether one of these ways of combining the same economic and stochastic primitives provides a markedly better description of observed or out-of-sample data.

Because we cannot test economic models independent of functional form assumptions for a finite number of observations, it is important to recognize that structural parameter estimates may well be sensitive to these assumptions. For example, if we were trying to estimate consumer surplus, we should be aware that it might make a tremendous difference that we assumed demand was linear, as opposed to constant elasticity. While this sensitivity to functional form can be viewed as a weakness, it also can be viewed as a strength. This is again because the "structure" in structural models forces researchers to grapple directly with the economic consequences of assumptions.

The "structure" in structural models also can affect statistical inferences about economic primitives. Here we have in mind the impact that a researcher's functional form choices can have on the size and power of hypothesis tests. When, as is usually the case, economic theory does not suggest functional forms or what other variables might be relevant in an application, researchers will be forced to make what may seem to be arbitrary choices. These choices can have a critical impact on inferences about parameters. For example, if a researcher wants to fail to reject a null hypothesis, then she should specify an extremely rich functional form with plenty of variables that are not part of the economic theory. Such a strategy will likely decrease the power of the statistical test. For instance, if a researcher would like to fail to reject the integrability conditions for her demand functions, she should include as many demographic variables as possible in order to soak up across-household variation in consumption. This will tend to reduce the apparent precision of the estimated price coefficients and make it difficult to reject the null hypothesis of integrability. Conversely, if she would like to reject integrability, then she should include few, if any, demographic controls. This would increase the apparent precision in the price coefficients and increase the likelihood of rejection for two reasons: (1) she has reduced the number of irrelevant variables; and, (2) the effect of price may be exaggerated by the omission of relevant variables that are correlated with prices.

This discussion underscores the delicate position empiricists are in when they attempt to "test" a particular parameter or theory. For this reason, structural modelers should experiment with and report how sensitive their inferences are to plausible changes in functional forms, or the inclusion and exclusion of variables not closely tied to economic theory.

Finally, we should emphasize that these putative advantages do not always mean structural models should be favored over nonstructural models. Indeed, there are many interesting applications where there is little or no useful economic theory to guide empirical work. We certainly do not believe this should stop the collection or description of data. When on the other hand there is a substantial body of economic theory to guide empirical work, researchers should take advantage of it. In some cases, there may be a

large body of economic theory on a particular topic, but that theory may have few implications for data. In this case, structural modelers can make important contributions by making it clear what is needed to link theory to data. By being clear about what in the theory the empirical researcher can estimate, it becomes easier for economists to improve existing theory.

The advantages of structural models of course do not all come for free. All economic theories contain assumptions that are not easily relaxed. While theorists sometimes have the luxury of being able to explore stylized models with simplifying assumptions, structural econometric modelers have to worry that when they use stylized or simplifying assumptions they will be dismissed as arbitrary, or worse: insensitive to the way the world "really works". This problem is compounded by the fact that economic data rarely come from controlled experimental settings. This means that structural econometric modelers often must recognize nonstandard ways in which nonexperimental data are generated and collected (e.g., aggregation and censoring). Such complications likely will force the structural modeler to simplify. The danger in all of these cases is that the structural model can then be seen as "too naive" to inform a sophisticated body of theory. We expect that readers can see this already in Example 1.

3.3. Evaluating structure – single equation models

The standard multiple linear regression model is a useful place to begin understanding issues that arise in evaluating and interpreting structural and nonstructural models. Consider the linear regression model, $y = \alpha + x\beta + \epsilon$. The mathematical structure of this model lends an aura of "structure" as to how y is related to x. What motivates this structure? A satisfactory answer to this question minimally must address why we are regressing y on x. From a statistical perspective, we can always regress y on x or x on y. The coefficients in these regressions can then be given statistical interpretations as the coefficients of best linear predictor functions. Issues of economic causality, however, are not resolved simply because a researcher puts y on the left-hand side and x on the right-hand side of a linear equation.

Economists estimating linear regression models usually invoke economic arguments to make a case that x causes y. Assuming that the researcher has made a convincing case, what should be made of the regression of y on x? Absent an economic model showing that y and x are linearly related, all one can say is that under certain conditions ordinary least squares regression will provide consistent estimates of a best linear predictor function. The regression does not necessarily deliver an estimate of how much the conditional mean of y changes with a one unit change in x, and certainly not the causal impact of a one-unit change in x on y.

Despite this, some researchers use regression coefficient signs to corroborate an economic model in the belief that multiple regressions "hold constant" other variables. For example, a researcher might develop a deterministic economic model that shows: "when x increases, y increases". This result then becomes the economic justification for using a regression of y on x and other variables to "test" the theory. One problem

with this approach is that unless the economic model delivers a linear conditional mean specification for y given x, the regression evidence about the sign of x need not match deterministic comparative statics predictions. In general, the empirical researcher must first use economics and statistics to demonstrate that the relevant economic quantity or comparative static effect can be identified using the available data and estimation technique. To see this point more clearly, consider the following example.

EXAMPLE 2. A microeconomist has cross-section data on a large number of comparable firms. The data consist of outputs, Q, in physical units, total costs, TC, and the firms' two input prices, p_K and p_L. The researcher's goal is to learn about the firms' (by assumption) common technology of production. The researcher decides to do this by estimating one of the following regression models:

$$\begin{aligned} &\text{Model 1:} \quad \ln \mathrm{TC}_i = \theta_0 + \theta_1 \ln Q_i + \theta_2 \ln p_{Ki} + \theta_3 \ln p_{Li} + \eta_i, \\ &\text{Model 2:} \quad \ln Q_i = \beta_0 + \beta_1 \ln \mathrm{TC}_i + \beta_2 \ln p_{Ki} + \beta_3 \ln p_{Li} + \epsilon_i. \end{aligned} \tag{2}$$

These specifications differ according to whether the natural logarithm of output or the natural logarithm of total costs is a dependent or independent variable.

Which specification makes better economic sense? In an informal poll of colleagues, we found most thought Model 1 was more sensible than Model 2. The logic most often given for preferring Model 1 is that it looks like a cost function regression. When asked how to interpret the parameters of this regression specification, most say that θ_1 is an estimate of the elasticity of total cost with respect to output. As such, it provides a measure of scale economies. Those who prefer the second equation seem to base their preference on an argument that total cost is more likely to be "exogenous". To them this means that OLS is more likely to deliver consistent estimates of production or cost parameters.

How might we go about deciding which specification is correct? A structural modeler answers this question by answering two prior questions: What economic and statistical assumptions justify each model? And, do these assumptions make sense for the application at hand? In Example 5 of Section 4, we show that Models 1 and 2 can be derived from competing plausible economic and stochastic assumptions. That is, under one set of economic and stochastic modeling assumptions, we can derive Model 1. Under another set of assumptions, we can do the same for Model 2. Without knowing the details of the firms and markets being studied, it is impossible to decide which set of assumptions is more appropriate.

How do researchers only interested in data description decide which specification is correct? They too must answer prior questions. But these questions only pertain to the goals of their statistical analysis. If, for example, their goal is prediction, then they would choose between Models 1 and 2 based on the variable they are trying to predict. They then would have to decide which right-hand side variables to use and how these variables would enter the prediction equation. Here, researchers have to worry

that if their goal is post-sample prediction, they may over-fit within sample by including too many variables. While statistical model selection criteria can help systematize the process of selecting variables, it is not always clear what one should make of the resulting model.

In some cases, researchers do not have a clear economic model or descriptive criterion in mind when they estimate a regression model such as Model 1 by ordinary least squares. In this case, what can be made of the coefficient estimates obtained from regressing y on the vector x? As discussed above, ordinary least squares delivers consistent estimates of the coefficients in the best linear predictor of y given x. But what information does the BLP$(y \mid x)$ provide about the joint distribution of y and x? In general, the BLP will differ from the more informative conditional expectation of y given x, $E(y \mid x)$, which is obtained from $f(x, y)$ as $\int yf(y \mid x)\,\mathrm{d}y$. Thus, $\theta_1 = \partial\,\mathrm{BLP}(y \mid x)/\partial x_1$ in Model 1 will not in general equal how much expected log total costs will increase if we increase log output by one unit (i.e., $\partial E(y \mid x)/\partial x_1$). Only under certain conditions on the joint density of y and x are the BLP function and the conditional expectation function the same. Despite this well-known general lack of equivalence between the BLP$(y \mid x)$ and $E(y \mid x)$, many descriptive studies treat linear regression slope coefficient estimates as if they were estimates of the derivative of $E(y \mid x)$ with respect to x. Occasionally, some studies adopt the position that while the best linear predictor differs from the conditional expectation, the signs of the coefficients of the BLP$(y \mid x)$ will be the same as those of $\partial E(y \mid x)/\partial x$ provided the signs of $\partial E(y \mid x)/\partial x$ do not change with x. Unfortunately, as White (1980) demonstrates, there is no reason to expect that this will be true in general.

When the conditional expectation of y is nonlinear in x, statistical theory tells us (under certain sampling assumptions) that a regression provides a best (minimum expected squared prediction error) linear approximation to the nonlinear conditional expectation function. It is perhaps this result that some place faith in when they attempt to use regressions to validate an economic comparative static result. However, absent knowledge from economics or statistics about the joint distribution of y and x, this approximation result may be of limited value. We do not, for example, know how good the linear approximation is. We do not know if x causes y or y causes x. In sum, BLP$(y \mid x)$ and $E(y \mid x)$ are simply descriptive statistical quantities.

By making economic and statistical assumptions, however, we can potentially learn something from the linear approximation. For example, if we had an economic theory that suggested that there was a negative causal relationship between y and z, then the bivariate regression slope coefficient's sign might tell us whether the evidence is consistent with the theory. But this may be a weak confirmation of the theory and it certainly does not provide us with a sense of the strength of the relationship if the conditional mean function, $E(y \mid z)$, is nonlinear in z.

Descriptive researchers (and structural modelers) also have to worry about whether they have collected all of the data needed to examine a particular prediction about a conditional mean. Consider, for example, the case where an economic theory delivers a prediction about the conditional mean of y given x_1 and x_2, $E(y \mid x_1, x_2)$, where

y, x_1 and x_2 are scalars. Suppose that y is a customer's demand for electricity during the day, x_1 is the price of electricity during the day, and x_2 is average temperature during the day. Economic theory predicts that electricity demand is decreasing in the daily price after controlling for the average daily temperature. However, if we do not include x_2 on the right-hand side when we regress y on x_1, then we obtain the best linear approximation to $E(y \mid x_1)$, not $E(y \mid x_1, x_2)$. The difference may be very important. For instance, the function $g(x_1) = E(y \mid x_1)$ may not depend on x_1, whereas the function $h(x_1, x_2) = E(y \mid x_1, x_2)$ may depend on both x_1 and x_2.

In the usual textbook analysis of omitted variables in a linear model, it is straightforward to establish when an omitted variable will cause bias and produce inconsistent estimates. When the conditional mean is nonlinear, and we proceed as if it is linear, the familiar reasoning is not as straightforward. In addition to the omitted variable, we have to worry that even if we had included the omitted variable, that $\partial E(y \mid x_1, x_2)/\partial x_1 \neq \partial \,\mathrm{BLP}(y \mid x_1, x_2)/\partial x_1$. Absent a theory that says that y is linearly related to x_1 and x_2, the effect of omitting a relevant regressor is much harder to evaluate. Specifically, suppose

$$y_t = g(x_{1t}, x_{2t}) + \epsilon_t = g(x_t) + \epsilon_t$$

and $E(\epsilon_t \mid x_t) = 0$ so that $E(y \mid x) = g(x_t)$ is a nonlinear function. Running an ordinary least squares regression of y_t on z_t, where z_t is a vector of known functions of x_{1t} and x_{2t}, yields a consistent estimate of β where β is defined as follows:

$$y_t = z_t\beta + \big[g(x_t) - z_t\beta\big] + \epsilon_t = z_t\beta + \eta_t.$$

The parameter β is the linear combination of the z_t's that best predicts the y_t's for the population. By construction $E(z_t\eta_t) = 0$, but the partial derivative of $z_t\beta$ with respect to x_1 could differ in both sign and magnitude from the partial derivative of the conditional mean, $\partial g(x_t)/\partial x_{1t}$ depending on how well $z_t\beta$ approximates $g(x_t)$.

3.4. Evaluating structure – simultaneous equation models

The original Cowles Commission econometricians paid particular attention to developing econometric models that could represent the concept of an "economic equilibrium". Indeed, the term "structural model" often is associated with econometric models that have multiple simultaneous equations, each of which describes economic behavior or is an identity. The term simultaneous emphasizes that the left-hand side variables also can appear as right-hand side variables in other equations. The term "reduced form" was introduced to describe an alternative representation of a simultaneous system – one in which the dependent variables were explicitly represented only as functions of the x's and unobservables.

To understand what is "structural" in simultaneous equations models, it is useful to begin with a standard linear supply and demand model.

EXAMPLE 3. In a standard linear demand and supply model, the demand curve gives the quantity that consumers would like to purchase at a given price, conditional on other variables that affect demand, and the supply curve gives how much firms are willing to sell at a given price, conditional on other supply shifters. Mathematically,

$$\begin{aligned} q_t^s &= \beta_{10} + \gamma_{12} p_t + \beta_{11} x_{1t} + \epsilon_{1t}, \\ p_t &= \beta_{20} + \gamma_{22} q_t^d + \beta_{22} x_{2t} + \epsilon_{2t}, \\ q_t^s &= q_t^d, \end{aligned} \tag{3}$$

or in matrix notation:

$$[\, q_t \quad p_t \,] \begin{bmatrix} 1 & -\gamma_{22} \\ -\gamma_{12} & 1 \end{bmatrix} - [\, 1 \quad x_{1t} \quad x_{2t} \,] \begin{bmatrix} \beta_{10} & \beta_{20} \\ \beta_{11} & 0 \\ 0 & \beta_{22} \end{bmatrix} = [\, \epsilon_{1t} \quad \epsilon_{2t} \,], \tag{4}$$

$$y_t' \Gamma - x_t' B = \epsilon_t', \tag{5}$$

where Γ and B are matrices containing the unknown parameters that characterize the behavior of consumers and producers, q_t is equilibrium quantity at time t, p_t is equilibrium price, y_t is a two-dimensional vector, ϵ_t is a two-dimensional vector of unobserved random variables, and the exogenous variables, x_t, consist of a constant term, a supply shifter x_{1t} (e.g., an input price) and a demand shifter x_{2t} (e.g., household income).

To find out what restrictions the system (3) imposes on the conditional distribution of y given x, we can first solve for the endogenous variables as a function of exogenous variables and error terms. Post-multiplying both sides of (5) by Γ^{-1}, and rearranging, gives the reduced form

$$y_t' = x_t' \Pi + v_t'. \tag{6}$$

The reduced form (6) shows that equilibrium prices and quantities are linear functions of both demand and cost shifters and both demand and cost errors.

From this perspective, the individual reduced forms for equilibrium price and quantity parallel the nonstructural descriptive linear regressions discussed in the previous subsection. Some researchers, however, over-extend this logic to claim that any regression of an endogenous variable on exogenous variables is a "reduced form" regression. Thus, they would for example label a regression of price on the supply shifters x_{2t} a "reduced form".

The critical issue again in any regression of a y on x is what do we make of the estimated coefficients? Returning to the reduced form system (6), to arrive at this representation we had to first assume that the structural equations characterizing aggregate demand and supply were linear. If we did not know they were linear, then we would not know that the reduced form (6) was linear. In short, the functional form of the structural model determines the functional form of the reduced form relationship between y_t and x_t.

The economic assumption that supply equals demand also is critical to the interpretation of Π. If, for example, price floors or ceilings prevented demand from equaling

supply, then we would not obtain a standard linear model even though the underlying demand and supply schedules were linear [see Quandt (1988)].

Even when we are assured that the reduced form has the linear form (6), we cannot interpret the Π and proceed to estimation without completing the specification of the demand and supply equations. To complete the structural model, for example, the researcher could specify the joint distribution of x and y, or alternatively, as is common in the literature, the conditional distribution of y given x. Still another approach is to sacrifice estimation efficiency by imposing less structure on the joint distribution. For example, estimation could proceed assuming the conditional moment restrictions

$$E(\epsilon_t \mid x_t) = 0 \quad \text{and} \quad E\left(\epsilon_t \epsilon_t' \mid x_t\right) = \Sigma. \tag{7}$$

From these conditional moment restrictions on ϵ_t, we can deduce

$$E(v_t \mid x_t) = 0, \quad \text{and} \quad E\left(v_t v_t' \mid x_t\right) = \Omega, \tag{8}$$

where

$$\Pi = B\Gamma^{-1}, \quad v_t' = \epsilon_t' \Gamma^{-1}, \quad \text{and} \quad \Omega = \left(\Gamma^{-1}\right)' \Sigma \Gamma^{-1}. \tag{9}$$

From (9), we see that Π and the variance–covariance matrix of the reduced form errors, Ω, provide information about the structural parameters in Γ. Without restrictions on the elements of Γ, B, and Σ, however, the only restriction on the conditional distribution of y_t given x_t implied by the linear simultaneous equation model is that the conditional mean of y_t is linear in x_t and the conditional covariance matrix of y_t is constant across observations.

To summarize, a *reduced form* model exists only to the extent that the researcher has derived it from a structural economic model. If the researcher is unwilling to assume functional forms for the supply and demand equations, then the conditional means of q_t and p_t will likely be nonlinear functions of x_t, the vector of the demand and supply shifters. In this case, although we can still perform linear regressions of q_t and p_t on x_t, these linear regressions are not reduced forms. Instead, these regressions will deliver consistent estimates of the parameters of the best linear predictors of the dependent variables given x_t. How these parameter estimates are related to the price elasticity of demand or supply or other causal effects is unknown. Additionally, as discussed earlier, unless the researcher is willing to place restrictions on the functional forms of the conditional means of q_t and p_t given x_t, it will be difficult to make even qualitative statements about the properties of $E(p_t \mid x_t)$ or $E(q_t \mid x_t)$.

Notice, it is economic theory that allows us to go beyond descriptive or statistical interpretations of linear regressions. If we assume stochastic linear supply and demand equations generate y_t, and impose the market-clearing conditions $q_t^s = q_t^d$, then the equations in (9) allow us *in principle* to recover estimates of economic parameters from Π and Ω. We emphasize *in principle* because unless the values of B, Γ, and Σ can be uniquely recovered from Π and Ω, the structural model (3) has limited empirical content. Although the structural parameters are not identified, the linearity of the structural

model implies that the conditional mean of y_t is linear in x_t and the conditional variance is constant.

One might wonder how we know that the structural model given in Equation (3) is generating the observed y_t. The answer is by now familiar: Only because economic theory tells us so! Economic theory tells us what elements of x_t belong in just the supply and just the demand equations. The same theory also resolves the problem of how to identify Γ, B, and Σ from the reduced form parameters Π and Ω. Absent restrictions from economic theory, there are many different simultaneous equations models that can give rise to the same reduced form parameters Π and Ω. These models may contain radically different restrictions on the structural coefficients and impose radically different restrictions on the behavior of economic agents, yet no amount of data will allow us to distinguish among them. For economic theory to be useful, it minimally must deliver enough restrictions on Γ, B, and Σ so that the empiricist can uniquely recover the remaining unrestricted elements of Γ, B, and Σ from estimates of Π and Ω. Thus, any defense of the researcher's identification restrictions can be seen as a defense of the researcher's economic theory. Without a clearly argued and convincing economic theory to justify the restrictions imposed, there is little reason to attempt a structural econometric model.

It is well known to economic theorists that without assumptions it is impossible to derive predictions about economic behavior. For example, consumers may have preference functions and producers access to technologies. However, unless we are willing to assume, for example, that consumers maximize utility subject to budget constraints and producers maximize profits subject to technological constraints, it is impossible to derive any results about how firms and consumers might respond to changes in the underlying economic environment. An empirical researcher faces this same limitation: without assumptions, it is impossible to derive empirical results. From a purely descriptive perspective, unless a researcher is willing to assume that the joint density of x and y satisfies certain conditions, he cannot consistently estimate underlying descriptive magnitudes, such as the $\text{BLP}(y \mid x)$ or the conditional density of y given x. Further, unless an empirical researcher is willing to make assumptions about the underlying economic environment and the form and distribution of unobservables, he cannot estimate economically meaningful magnitudes from the resulting econometric model. So it is only the combination of economic and statistical assumptions that allow conclusions about economic magnitudes to be drawn from the results of an econometric modeling exercise.

Econometrics texts are fond of emphasizing the importance of exclusion restrictions for identification – an exogenous variable excluded from the equation of interest. We would like to emphasize that identification also requires inclusion restrictions – this exogenous variable must also be included in at least one equation of the structural model.

This distinction is particularly important because applied researchers are typically unwilling or unable to specify all the equations in their simultaneous equations system. This incompleteness in the econometric model reflects an incompleteness in the eco-

nomic model. This incompleteness can and should raise doubts about the validity of instruments. To see why, suppose economic theory delivers the following linear simultaneous equations model

$$y_1 = \beta y_2 + x_1\gamma + \epsilon_1,$$
$$y_2 = x_1\pi_{21} + \epsilon_2, \tag{10}$$

where the ϵ's are independently and identically distributed (i.i.d.) contemporaneously correlated errors and x_1 is a variable that is uncorrelated with ϵ_1 and ϵ_2. Suppose that a researcher is interested in estimating the structural parameters β and γ in the first equation. As it stands, these parameters are not identified. The problem is that we are missing an instrument for y_2.

What to do? One approach is to revisit the economic theory in an effort to understand where additional instruments might come from. An alternative approach that is all too common is the recommendation: "Look for an exogenous variable that is uncorrelated with the ϵ's but at the same time correlated with the right-hand side endogenous variable y_2". While these two approaches are not necessarily incompatible, the second approach does not seem to involve any economics. (This should sound a warning bell!) All one needs to find is a variable that meets a statistical criterion. In some instances, researchers can do this by searching their data sets for variables that might reasonably be viewed as satisfying this criterion.

The following suggests how a researcher might run into problems using the statistical approach: "Look for an exogenous variable that is uncorrelated with the ϵ's but at the same time correlated with the right-hand side endogenous variable y_2". Consider first the extreme case where we decide to create a computer-generated instrument for y_2 that satisfies this criterion. That is, imagine we construct an instrumental variable, x_2, as the sum of x_1 plus a computer-generated independent identically distributed random error. This new variable satisfies the statistical criteria to be a valid instrument: it is uncorrelated with the structural errors and yet correlated with y_2. Thus, it would appear that we can always identify the coefficients in the first equation as long as we have at least one exogenous variable and a good random number generator.

What is amiss here is that identification also hinges on showing that x_2 belongs in the second equation. A statistical test cannot unambiguously resolve this question (especially when x_1 and x_2 are highly correlated). However, both economics and common sense tell us that x_2 does not belong in the reduced form. Put another way, they tell us that x_2 does not belong in the structural or reduced form model – the population value of π_{22}, the coefficient associated with x_2 in the reduced form, is *zero*! Nevertheless, in finite samples we could conclude π_{22} is nonzero (and perhaps statistically so).

To understand formally why this estimation strategy fails to produce consistent estimates of β and γ, consider the instrumental variables estimator for these two parameters. This estimator uses the instruments $(x_1, x_2)'$:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T} y_{2t}x_{1t} & \sum_{t=1}^{T} x_{1t}^2 \\ \sum_{t=1}^{T} y_{2t}x_{2t} & \sum_{t=1}^{T} x_{1t}x_{2t} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{T} x_{1t}y_{1t} \\ \sum_{t=1}^{T} x_{2t}y_{1t} \end{bmatrix}.$$

A necessary condition for the consistency of this instrumental variables estimator is that the matrix

$$\frac{1}{T}\begin{bmatrix} \sum_{t=1}^{T} y_{2t}x_{1t} & \sum_{t=1}^{T} x_{1t}^2 \\ \sum_{t=1}^{T} y_{2t}x_{1t} & \sum_{t=1}^{T} x_{1t}x_{2t} \end{bmatrix}$$

converges in probability to a finite nonsingular matrix. Assume that

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} x_{1t}^2 = M_2.$$

Because $x_{2t} = x_{1t} + \eta_t$ and η_t is distributed independently of ϵ_{1t}, ϵ_{2t}, and x_{1t}, the probability limit of this matrix is equal to

$$\begin{bmatrix} M_2\pi_{21} & M_2 \\ M_2\pi_{21} & M_2 \end{bmatrix}, \tag{11}$$

which is a singular matrix. This result follows from substituting $x_{1t} + \eta_t$ for x_{2t} and $x_{1t}\pi_{21} + \epsilon_{2t}$ for y_{2t} and then applying the appropriate laws of large numbers to each element of the matrix. The singularity of (11) is just another way of saying that the rank condition for identification of the first equation of the structural model fails.

At first, this example may seem extreme. No economist would use a random number generator to create instruments – but this is our point! The researcher is informed not to do this by economics. In practice, a researcher will never know whether a specific instrument is valid. For example, our students sometimes insist that more clever choices for instruments would work. After some thought, many suggest that setting $x_2 = x_1^2$ would work. Their logic is that if x_1 is independent of the errors, so must x_1^2. Following the derivations above, and assuming that $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} x_{1t}^3 = M_3$, a finite, positive constant, we again obtain a singular matrix similar to (11), implying that this x_2 is also an invalid instrument for the same reason – it does not enter into the reduced form.

The value of economic theory is that it provides a defense for why the reduced form coefficient on a prospective instrument is not zero, i.e., the instrument is included in at least one equation of the structural model. The statistical advice that led to computer-generated instruments and x_1^2 does not do this.[2]

Some might argue that our example above ignores the fact that in most economic applications, one can find exogenous economic variables that satisfy our statistical criterion. The argument then goes on to claim that because these variables are economically related, we do not need a complete simultaneous equations model. The following example discusses this possibility.

[2] An element of x_t is a valid instrument in linear simultaneous equations model if it satisfies the conditional moment restrictions (7), $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} x_t x_t' = Q$, where Q is a positive definite matrix, and it enters at least one of the equations of the structural model. Our computer generated instrument fails this last requirement.

EXAMPLE 4. Consider a researcher who has data on the prices firms charge in different geographic markets, p_i, the number of potential demanders (population) in market i, POP_i, and whether or not the firm faces competition, COMP_i. The researcher seeks to measure the "effect" of competition on prices by regressing price on market size, as measured by the number of potential demanders and the competition dummy. That is, he estimates the regression

$$p_i = \text{POP}_i\,\theta_1 + \text{COMP}_i\,\theta_2 + \epsilon_i. \tag{12}$$

Without an underlying economic model, the OLS estimate of θ_2 on COMP_i provides an estimate of the coefficient in the best linear predictor of how prices change with the presence of competition.

The researcher might, however, claim that Equation (12) has a structural economic interpretation – namely that θ_2 measures by how much prices would change if we could introduce competition. One problem with this interpretation is that it is unlikely that the presence of competition is determined independently of price. (See Section 10.) In most entry models, competitors' decisions to enter a market are simultaneously determined with prices and quantities. In such cases, if the researcher does not observe critical demand or supply variables, then OLS will deliver inconsistent estimates of θ_2.

One possible solution to this problem is to find an instrumental variable for the presence of competitors. Suppose that the researcher claims that the average income of residents in the market, Y_i, is such an instrument. This claim might be justified by statements to the effect that the instrument is clearly correlated with the presence of competitors, as an increase in average income, holding population fixed, will increase demand. The researcher also might assert that average income is determined independently of demand for the good and thus will be uncorrelated with the error ϵ_i in Equation (12).

Does this make average income a valid instrument? Our answer is that the researcher has yet to make a case. All the researcher has done is provide a statistical rationale for the use of Y_i as an instrument exactly analogous to the argument used to justify the computer-generated instrument in Example 3. To be completely convincing, the researcher must do two more things. First, the researcher has to explain why it makes sense to *exclude* average income from Equation (12). To do this, the researcher will have to provide a more complete economic justification for Equation (12). What type of equilibrium relationship does Equation (12) represent? Why is the demand variable POP_i in this equation, but not average income, which also might be considered a demand variable? Second, the researcher also will have to make a case that Y_i enters the reduced form for COMP_i with a nonzero coefficient, or else the rank condition for identification will fail by the logic presented in Example 3. This means to be a valid instrument it must enter some other equation of the structural model. The researcher will have to be clearer about the form of the complete system of equations determining prices and the presence of competitors. This will also require the researcher to spell out the economic model underlying the simultaneous system of equations.

This next example reiterates our point that the results of a structural modeling exercise are only as credible as the economic theory underlying it. One can always impose inclusion and exclusion restrictions, but the resulting simultaneous equations model need not have a clear interpretation.

EXAMPLE 5. The 1960s' and 1970s' IO literature contains many studies that regressed firm or industry profit rates ("performance") on market concentration measures ("market structure"). In the late 1960s and early 1970s, many IO economists observed that while concentration could increase profits, there could be the reverse causation: high (low) profits would induce entry (exit). This led some to estimate linear simultaneous equations models of the form:

$$\begin{aligned} PROFIT &= \beta_0 + \beta_1 CONC + x_1\beta_2 + \epsilon_1, \\ CONC &= \alpha_0 + \alpha_1 PROFIT + x_2\alpha_2 + \epsilon_2, \end{aligned} \tag{13}$$

where *PROFIT* measures industry profitability, *CONC* measures industry concentration, the ϵ's are errors and the α's and β's are parameters to be estimated. Particular attention was paid to estimating the effect of simultaneity bias on the signs and magnitudes of α_1 and β_1.

Debates about the merits of these models often centered on what variables should be included or excluded from each equation. What proved unsatisfactory about these debates was that there were no clear answers. Put another way, although these were called "structural" models of performance and market concentration, there was no one theoretical model that provided a specific economic interpretation of α_1 and β_1. Thus, even though instrumental variable methods might deliver consistent estimates of α_1 and β_1, it was never very clear what these estimates told us about the underlying theories.

To understand why we would not call this a structural model (even though it looks like a "structural" model in the sense of having multiple endogenous variables in a single equation), consider these questions: How do we know the first equation is a behavioral relation describing how industry profitability responds to industry concentration? And: How do we know the second equation describes the way firm profitability responds to industry concentration? The population values of β_1 and α_1, the parameters that characterize how *PROFIT* responds to *CONC* and how *CONC* responds to *PROFIT*, depend crucially on which elements of x_t are included and excluded from each equation of the structural model. Unless we have an economic theory telling us which elements of x_t do and do not belong in each behavioral relation, which equation we designate as the "profit equation" and which equation we designate as a "concentration equation" is completely arbitrary. For example, we can re-write the "profit equation" in (13) as a "concentration equation",

$$\begin{aligned} CONC &= -\frac{\beta_0}{\beta_1} + \frac{1}{\beta_1} PROFIT - x_1\frac{\beta_2}{\beta_1} - \frac{1}{\beta_1}\epsilon_1 \\ &= \theta_0 + \theta_1 PROFIT + x_1\theta_3 + \eta. \end{aligned}$$

What is the difference between this "concentration equation" and the one in (13)? Because they obviously have the same left-hand side variable, the answer is that they differ in what is included on the right-hand side. One has x_1 and the other has x_2. Only economic theory can tell us which "concentration equation" is correct. That is, only economic theory can tell us what agent's behavior, or group of agents' behaviors, is characterized by a structural equation. For example, we might try to justify the profit equation in (13) as representing the profit-maximizing behavior of all firms in the industry for each level of industry concentration and conditioning variables. It is this same theory that tells us the conditioning variables in the profit equation are the x_1's and not the x_2's. Thus, economic theory also delivers the inclusion and exclusion restrictions that allow us to interpret the equations of structural econometric models.

In his criticism of large-scale macroeconometric models, Sims (1980) referred to many of the restrictions used to identify macro models as "incredible". He observed: "the extent to which the distinctions among equations in large macromodels are normalizations, rather than truly structural distinctions, has not received much emphasis" [Sims (1980, p. 3)]. By truly structural distinctions, Sims meant exclusion and other functional form restrictions derived from economic theory. This same criticism clearly applies to structural modeling of the relationship between profits and concentration. As we describe in later sections, the lack of satisfactory answers to such questions is what led some empirical IO economists to look more closely at what economic theory had to say about firm profitability and market concentration.

3.5. The role of nonexperimental data in structural modeling

Virtually all data used in empirical economic research comes from nonexperimental settings. The use of nonexperimental data can raise significant additional modeling issues for descriptive and structural modelers. Consider a researcher who wants to describe the relationship between firms' prices and the number of competitors. Suppose that the data under consideration come from markets where firms face a price cap. The most general approach to describing this relationship would be to estimate flexibly the joint distribution of prices and competitors. Provided the price cap is binding in some markets, the researcher would obtain a density that has a spike at the price cap.

Instead of flexibly estimating the joint distribution of prices and competitors, the researcher could instead use a regression to describe the relationship. As we argued earlier, OLS will deliver consistent estimates of the best linear predictor function. Suppose that absent the cap economic theory implied that the conditional mean of prices given the number of competitors (x) was linear in x. How does the presence of the cap affect the estimation of the coefficients of the conditional mean function? The answer is that the cap truncates the joint distribution and thereby alters the conditional mean of y given x. Thus, the researcher will need to model this truncation if he is to recover a consistent estimate of the coefficients in the conditional mean function.

Although similar statistical sampling issues can arise in structural models, a structural econometric modeler would view the presence of a price cap as more than a statistical nuisance. Rather, the cap is something that needs to be accounted for in the modeling of firm behavior and the unobservables.

To illustrate how structural models can account for nonexperimental data, let us return to the demand and supply model for prices and quantities. Suppose the researcher observes price, the total quantity that consumers demand at that price, and consumer income (x_1). Suppose also that the researcher has estimated the regression

$$q_t^s = \beta_0 + \beta_1 p_t + \beta_2 x_{1t} + \epsilon_{1t}$$

by OLS. For the researcher to be able to assert that they have estimated a demand curve, as opposed to a descriptive best linear predictor, they must be able to argue that price and income are uncorrelated with the error. When is this likely the case? In principle, it would be the case if the researcher could perform experiments where they faced all consumers with a random series of prices. The same experiment also could be used to estimate a supply equation using OLS, provided the researcher observed the quantity supplied at the randomly chosen price.

The key feature of the experiment that makes it possible to estimate both the demand and supply equations by OLS is that the researcher observes both the quantity demanded and the quantity supplied at each randomly chosen price. In general, the quantity demanded *will not equal* the quantity supplied at a randomly chosen price. In other words, the third equation in the demand and supply system (3) does not hold.

How do equilibrium models of price determination compare then to experimental models? One way to view nonexperimental data is that it came from a grand experiment. Imagine that in this grander experiment, the experimentalist had collected data for a vast range of randomly selected prices, incomes and input prices. Imagine now someone else extracts from the experimentalist's data only those observations in which the experimenter's randomly chosen prices, incomes and input prices resulted in the quantity supplied equaling the quantity demanded. This nonrandom sample selection would yield a data set with significantly less information and, more importantly, nonrandom prices. Thus, even though the original data came from a random experiment, the data selection process will cause OLS to no longer deliver consistent estimates of the supply and demand parameters. On the other hand, if the researcher were to apply instrumental variable techniques appropriate for a structural simultaneous equations model that (correctly) imposed the market-clearing equation (3), they would obtain consistent estimates.

Our general point here is that structural models are valuable in nonexperimental contexts because they force the researcher to grapple directly with nonexperimental aspects of data. Consider again the demand and supply model above. How did we know it was appropriate to impose $q^s = q^d$? The answer came not from a statistical model of the nonrandomness, but from our economic perspective on the nonexperimental data – we assumed that the data came from markets where there are no price floors or ceilings. Had

there been price floors or ceilings, this would change the third equation in our econometric model. For example, with binding price ceilings, we might assume that the quantity we observe is the quantity supplied. (With a binding ceiling, quantity demanded exceeds supply, but we typically would not know by how much.) Our econometric model now would have to account for this selection of quantities. A variety of such "disequilibrium" demand and supply models exist and are reviewed in Maddala (1983).

4. A framework for structural econometric models in IO

Having described differences between descriptive and structural models, we now provide a framework for building and evaluating structural econometric models. While in principle it would seem easy for empiricists to recast an economic model as an econometric model, this has not proven true in practice. The process of combining economic and statistical models is by no means formulaic. As we have indicated earlier, the process of building a tractable econometric model that respects the institutions being modeled often involves difficult trade-offs. In the remaining sections we will use the framework to illustrate the progress of structural modeling in IO.

Structural modeling, and the elements of our framework, are not new to IO or most applied fields in economics. More than fifty years ago, Trygve Haavelmo and economists at the Cowles Foundation began combining models of individual agent behavior with stochastic specifications describing what the econometrician does not know:

> The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier. ... So far, the common procedure has been to first construct an economic theory involving exact functional relationships, then to compare this theory with some actual measurements, and finally "to judge" whether the correspondence is "good" or "bad". Tools of statistical inference have been introduced, in some degree, to support such judgment ... [Haavelmo (1944, p. iii)]

While the general principle of combining economic models with stochastic specifications has been around for some time, each field of economics has had to confront its own problems of how best to combine models with data. Often the desire to have a simple, well-defined probability model of the endogenous variables forces compromises. Early on, Hood and Koopmans described the challenge facing empirical economists as:

> In reality, unobserved random variables need to be introduced to represent "shocks" in behavior relations (i.e., the aggregate effects on economic decisions of numerous variables that are not separately observed) and "errors" of measurement. The choice of assumptions as to the distribution of these random variables is further complicated by the fact that the behavior equations in question are often aggregated over firms or individuals. The implications of this fact are insufficiently explored so far. [Hood and Koopmans (1953, p. xv)]

Following in this tradition, we describe a procedure for structural economic modeling that contains three basic steps. The first step is to formulate a well-defined economic model of the environment under consideration. The second step involves adding a sufficient number of stochastic unobservables to the economic model, so that its solution produces a joint density for all observables that has positive support on all possible realizations of these variables. The final step involves verifying the adequacy of the resulting structural econometric model as a description of the observed data.

4.1. The economic model

The first main component of a structural model is a complete specification of the equations describing economic behavior, what we call the economic model. Almost all economic models in IO have the following five components:

1. A description of the economic environment, including:
   (a) the extent of the market and its institutions;
   (b) the economic actors; and
   (c) the information available to each actor.
2. A list of primitives, including:
   (a) technologies (e.g., production sets);
   (b) preferences (e.g., utility functions); and
   (c) endowments (e.g., assets).
3. Variables exogenous to agents and the economic environment, including:
   (a) constraints on agents' behavior; and
   (b) variables outside the model that alter the behavior of economic agents.
4. The decision variables, time horizons and objective functions of agents, such as:
   (a) utility maximization by consumers and quantity demanded; and
   (b) profit maximization by firms and quantity supplied.
5. An equilibrium solution concept, such as:
   (a) Walrasian equilibrium with price-taking behavior by consumers; and
   (b) Nash equilibrium with strategic quantity or price selection by firms.

While the rigor of mathematics forces theorists to be clear about these components when they build an economic model, structural econometric models differ considerably in the extent to which they spell out these components. Our later discussions will illustrate the value of trying to make these components clear. In particular, we will focus attention on component 5, the equilibrium solution concept, because this is the most critical and specific to IO models.

4.2. The stochastic model

The next step in structural modeling is unique to empirical research. It receives much less attention than it deserves. This step is the process by which one transforms a deterministic (or stochastic) economic model into an econometric model. An econometric

model is distinct from an economic model in that it includes unobservables that account for the fact that the economic model does not perfectly fit observed data. Our main point is that the process of introducing errors should not be arbitrary. Both the source and properties of these errors can have a critical impact on the distribution of the observed endogenous variables and estimation.

The four principal ways in which a researcher can introduce stochastic components into a deterministic economic model are:

1. researcher uncertainty about the economic environment;
2. agent uncertainty about the economic environment;
3. optimization errors on the part of economic agents; and
4. measurement errors in observed variables.

This subsection emphasizes how these stochastic specifications differ, and in particular how they can affect the manner by which the researcher goes about estimating structural parameters.

4.2.1. Unobserved heterogeneity and agent uncertainty

A researcher's uncertainty about the economic environment can take a variety of forms. These different forms can have dramatically different implications for identification and estimation. For this reason it is critical for structural modelers to explain where error terms come from and whose uncertainty they represent. A critical distinction that needs to be drawn in almost every instance is: Is the uncertainty being introduced shared by the economic actors and econometrician?

A common assumption is that the researcher knows much less about the economic environment than the economic agents. In this case, the economic agents base their decisions on information that the researcher can only include in an error term. For example, if the researcher did not observe auction bidders' private information about an object, then the researcher would be forced to model how this unobservable information impacted bids. In general, we refer to a situation where agents' decisions depend on something the economist does not observe as a case of unobserved heterogeneity.

Of course researchers and economic agents can share uncertainty about the economic environment under study. For instance, the bidder may know their value for an object, but not the private values of the other bidders. In each of these cases, the firm or agent is presumed to know the distribution of uncertainty and make decisions that optimize the expected value of an objective function.

It might seem that because the econometrician is ignorant in both cases that unobserved heterogeneity and agent uncertainty are two sides of the same coin – they both rationalize introducing error terms in a structural model. The distinction, however, often is important for determining which estimation procedure is appropriate. To underscore this point, we now return to the two models described in (2). We shall show that, de-

pending on our assumptions about the source of the errors, it may be appropriate to regress ln TC on ln Q and other controls, or ln Q on ln TC and these same controls.

EXAMPLE 6. Imagine that we have cross-section data on comparable firms consisting of output, Q, total costs, TC, and input prices, p_K and p_L. Our goal is to estimate α and β in the Cobb–Douglas production function

$$Q_i = A_i L_i^{\alpha} K_i^{\beta}$$

consistently, where the subscript i denotes the ith firm. Because we do not have labor and capital information, we need to derive a relationship between total costs and output. There are many possible ways of doing this, each depending on what additional assumptions we make about the economic environment in which firms make their decisions.

Suppose, for example, that the firms are in a regulated industry, and have different A_i. For the purposes of exposition, assume that demand is completely inelastic. Consider now the case of pure unobserved heterogeneity (Type 1 shocks), where A_i is observed by the firm and the regulator, but not the econometrician. For simplicity, assume the A_i are i.i.d. positive random variables. Firm profits equal:

$$\pi(p_i, K_i, L_i) = p_i A_i L_i^{\alpha} K_i^{\beta} - p_{Ki} K_i - p_{Li} L_i.$$

Suppose that the regulator chooses p_i, the price of firm i's output first, and the firm then chooses K_i and L_i. Because demand is inelastic, a regulator interested in maximizing consumer welfare will set the firm's output price equal to the minimum average cost of producing Q_i. At this price, p_i^r, the firm chooses its inputs to minimize costs given the regulator's price and Q_i. That is, the firm maximizes

$$\pi(K_i, L_i) = p_i^r A_i L_i^{\alpha} K_i^{\beta} - p_{Ki} K_i - p_{Li} L_i.$$

Solving the firm's profit-maximizing problem, yields the total cost function:

$$\mathrm{TC}_i = C_0 p_{Ki}^{\gamma} p_{Li}^{1-\gamma} Q_i^{\delta} A_i^{-\delta}, \tag{14}$$

relating firm i's observed total cost data to its output. In this equation, $\delta = 1/(\alpha + \beta)$ and $\gamma = \beta/(\alpha + \beta)$. We can transform this total cost function into a regression equation using natural logarithms:

$$\ln \mathrm{TC}_i = \ln C_0 + \gamma \ln p_{Ki} + (1 - \gamma) \ln p_{Li} + \delta \ln Q_i - \delta \ln A_i. \tag{15}$$

While this equation holds exactly for the firm, the researcher does not observe the A_i. The researcher thus must treat the efficiency differences as unobservable in this logarithm of total cost regression:

$$\ln \mathrm{TC}_i = C_1 + \gamma \ln p_{Ki} + (1 - \gamma) \ln p_{Li} + \delta \ln Q_i - \delta \ln u_i. \tag{16}$$

This regression equation contains the mean zero error term

$$\ln u_i = \ln A_i - E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln Q_i].$$

The new constant term $C_1 = \ln C_0 + E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln Q_i]$ absorbs the nonzero conditional mean of the efficiency differences. Because the A_i are i.i.d., the conditional expectation reduces to the unconditional expectation, $E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln Q_i] = E[\ln A_i]$.

To summarize, we have derived a regression equation that is linear in functions of the (regulated) firm's production parameters. The relationship includes an error term that represents the firms' unobserved productive efficiencies. This error term explains why, at the same output level and input prices, firms could have different total costs. What is left to explain, is how a researcher would estimate the production parameters. This is a nontrivial issue in general. Here it is possible to argue that under fairly weak assumptions on the distribution of the u_i we can use ordinary least squares (OLS) to recover the production parameters. Note that OLS is appropriate because we have assumed that the regulator (and not the firm) picks price to recover the firm's minimum production cost to serve output Q_i. Put another way, OLS works because the unobserved heterogeneity in firms' production efficiencies is unrelated to the left-hand side regressors: firm output (which is inelastically demanded) and input prices (inputs are elastically supplied).

Now suppose that we observe the same data, but that the firm, like the econometrician, does not know its productive efficiency, A_i. This assumption leads to a different estimation strategy. In this case, the firm now must make its input decisions before it knows A_i. As long as the firm cannot undo this choice once A_i is realized, the firm maximizes expected profits taking into account the distribution of A_i. Now firm i's expected profit function is:

$$E\big[\pi(p_i, L_i, K_i)\big] = E\big[p_i A_i L_i^{\alpha} K_i^{\beta}\big] - p_{Ki} K_i - p_{Li} L_i. \tag{17}$$

We should note here that the expectation operator represents the firm's expectation.

Assume that the regulator again chooses p_i; the firm then chooses K_i and L_i. For simplicity, suppose that the regulator and the firm have the same uncertainty about the firm's productive efficiency. Suppose additionally that the regulator sets price, p_i^{er}, such that the firm earns zero profits in expectation. The firm then maximizes:

$$E\big[\pi\big(p_i^{\mathrm{er}} K_i, L_i\big)\big] = p_i^{\mathrm{er}} E\big[A_i L_i^{\alpha} K_i^{\beta}\big] - p_{Ki} K_i - p_{Li} L_i. \tag{18}$$

The first-order conditions for expected profit maximization imply

$$L_i = \left[\frac{\alpha p_{Ki}}{\beta p_{Li}}\right] K_i. \tag{19}$$

Observed total costs therefore equal

$$\mathrm{TC}_i = \frac{\alpha + \beta}{\beta} p_{Ki} K_i \tag{20}$$

and do not depend on the firm's (random) efficiency parameter A_i. Substituting these two expressions into the production function, we obtain an equation relating the observed (random) output Q_i^a to the firm's input prices and total costs

$$Q_i^a = D_0 \,\mathrm{TC}_i^{\alpha+\beta}\, p_{Ki}^{-\beta} p_{Li}^{-\alpha} A_i. \tag{21}$$

From both the firms' and the econometrician's perspective, the sole source of randomness here is the efficiency parameter A_i. Taking natural logarithms of both sides we obtain a regression equation that is linear in the production parameters

$$\ln Q_i^a = \ln D_0 + (\alpha + \beta) \ln \mathrm{TC}_i - \beta \ln p_{Ki} - \alpha \ln p_{Li} + \ln A_i. \tag{22}$$

This equation exactly explains firm i's realized production Q_i^a (which differs from the inelastically demanded quantity Q_i). Neither the firms nor the econometrician knows the A_i *ex ante*. Because the researcher also does not observe the efficiencies *ex post*, she must treat the efficiencies as random errors. She thus estimates the regression

$$\ln Q_i = D_1 + (\alpha + \beta) \ln \mathrm{TC}_i - \beta \ln p_{Ki} - \alpha \ln p_{Li} + \eta_i, \tag{23}$$

where $\eta_i = \ln A_i - E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln \mathrm{TC}_i]$. The constant term $D_1 = \ln D_0 + E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln \mathrm{TC}_i]$ absorbs the nonzero conditional mean of the efficiency differences. Once again, using the i.i.d. assumption on the A_i, the conditional expectation on $\ln A_i$ simplifies to the unconditional expectation. We can now use OLS to estimate the production parameters because by assumption the uncertainty in production is realized after the firm makes its production decision and is unrelated to total costs and input prices.

This example illustrates how the structural model's economic and stochastic assumptions can have a critical bearing on the consistency of a particular estimation strategy. Under one set of economic and stochastic assumptions, OLS applied to Equation (16) yields consistent estimates of the parameters of the firm's production function; under another set, we swap the dependent variable for one independent variable. Both models assumed (expected) profit-maximizing firms and (expected) welfare-maximizing regulators. In the first case, the stochastic shock represented only the researcher's ignorance about the productivity of firms. In the second, case, it represented uncertainty on the part of the firm, the regulator and the researcher about the productivity of the firm.

This example illustrates our initial point that a researcher should decide between models based upon how well their economic and stochastic assumptions match the environment in which the researcher's data were generated. Because no economic model is perfect in practice, the researcher often will be left choosing among imperfect assumptions and models. No statistical test can tell which model is correct. In later sections, we will discuss in more detail how a researcher might go about choosing among competing models.

4.2.2. Optimization errors

The third type of error listed above, optimization error, has received the least attention from structural modelers. In part, optimization errors have received less attention because there are few formal decision-theoretic models of optimization errors. The errors we have in mind are best illustrated by the behavior of economic agents in experiments. Experimental subjects often make errors, even when faced with relatively simple tasks.

Experimentalists' interpretations of these errors has been the source of considerable debate (e.g., see Camerer's (1995) survey). Here, we adopt a narrow view of what optimization error means so that we can illustrate the potential significance of such errors for structural models.

EXAMPLE 7. This example narrowly interprets optimization errors as the failure of agents' decisions to satisfy exactly first-order necessary conditions for optimal decisions. We are silent here on what causes this failure, and focus instead on its consequences. As an example, consider the standard consumer demand problem with unobserved heterogeneity in the utility function:

$$\min_{\lambda \geqslant 0}\Big[\max_{x \geqslant 0} U(x, \eta) + \lambda(M - p'x)\Big], \tag{24}$$

where x is an n-dimensional vector of consumption goods, p is the vector of prices, and M is the consumer's total budget. The vector η represents elements of individual tastes that the researcher does not observe. The normal first-order condition for x_i, assuming η is known is

$$\frac{\partial U}{\partial x_i}(x_i, \eta_i) - \lambda_i p_i = 0. \tag{25}$$

These equations yield the $i = 1, \ldots, n$ Marshallian demands, $x_i(p, M, \eta)$. In this case, the agent's first-order conditions are assumed to hold with probability one, so that for all realizations of η, all of the integrability conditions hold for the $x_i(p, M, \eta)$.

Now suppose that we introduce an additional source of error into the agent's demands. Although there are several ways to introduce error, imagine the errors do not impact the consumer's budget constraint (i.e., we still have $M = \sum_{i=1}^{n} p_i x_i$), but do impact the first-order conditions (25). Specifically, suppose

$$\frac{\partial U}{\partial x_i}(x, \eta) - \lambda p_i \nu_i = 0. \tag{26}$$

The researcher does not observe the ν_i, and thus treats them as random variables. Suppose for convenience that the researcher believes these errors have positive support and a mean of one in the population, so that on average the first-order conditions are correct.

How do the ν_i impact agents' decisions? If we solve the first-order conditions, and use the budget constraint, we obtain the Marshallian demand functions $x_i(p, M, \eta, \nu)$. Although the "demand curves" that result from this process satisfy homogeneity of degree zero in prices and total expenditure, they do not necessarily have a negative semi-definite Slutsky matrix for all realizations of the vector ν.

The next example shows how optimization errors can be used to rationalize why two seemingly identical consumers who face the same prices may purchase different amounts of x and y.

EXAMPLE 8. Imagine that we have demand data from a cross-section of similar consumers, all of whom have the same budget M, which they spend on two goods x and y. How should we model the differences in their consumption? One possible modeling strategy would be to say consumers have different preferences. Another would be to assume consumers have the same preference function, but that they make optimization errors when they make decisions.

Suppose each consumer has the utility function $U(x, y) = x^a y^b$ and that the first-order conditions have the form given in (26). Solving the first-order conditions, yields

$$\frac{a}{x} = \lambda p_x \nu_{xi}, \qquad \frac{b}{y} = \lambda p_y \nu_{yi}, \qquad p_x x + p_y y = M, \tag{27}$$

where λ is the Lagrange multiplier associated with the budget constraint and ν_{xi} and ν_{yi} are positive random variables representing optimization errors for consumer i. Further algebra yields

$$\lambda = \frac{\alpha_i + \beta_i}{M} \quad \text{with } \alpha_i = \frac{a}{\nu_{xi}} \text{ and } \beta_i = \frac{b}{\nu_{yi}}, \tag{28}$$

$$x = \frac{\alpha_i}{\alpha_i + \beta_i} \frac{M}{p_x} \quad \text{and} \quad y = \frac{\beta_i}{\alpha_i + \beta_i} \frac{M}{p_y}. \tag{29}$$

These demand functions look exactly like what we would get if there were no optimization error, and we had instead started with the Cobb–Douglas utility function $U(x, y) = x^{\alpha_i} y^{\beta_i}$. In other words, if we had started the modeling exercise by assuming that consumers did not make optimization errors, but instead had Cobb–Douglas preferences with heterogeneous utility parameters, we would have obtained an observationally equivalent demand model. The only way we might be able to distinguish between the two views would be to have data on consumers' choices across different purchase occasions. In this case, if consumers' tastes were time invariant, but their optimization errors varied intertemporally, we could in principle distinguish between optimization error and unobserved heterogeneity in tastes.

Optimization errors also can reduce the perceived rationality of agents' behavior. The following example shows that the way in which optimization errors are introduced can affect the extent to which firms are observed to be optimizing.

EXAMPLE 9. Consider a set of firms that have the common production function $Q = L^\alpha K^\beta$. Suppose each firm makes optimization errors when it attempts to minimize production costs. Specifically, assume that the factor demand functions are generated by solving the following three equations:

$$p_L \nu_L = \lambda \alpha K^\beta L^{\alpha-1}, \quad p_K \nu_K = \lambda \beta K^{\beta-1} L^\alpha, \quad \text{and} \quad Q = K^\beta L^\alpha, \tag{30}$$

where λ is the Lagrange multiplier associated with the constraint that the firm produces using the production function, and ν_{Li} and ν_{Ki} are unit mean, positive random variables representing optimization errors for firm i. Solving these three equations yields

following two factor demands:

$$L = Q^{\frac{1}{\alpha+\beta}} \left[\frac{p_K}{p_L}\right]^{\frac{\beta}{\alpha+\beta}} \left[\frac{\beta v_K}{\alpha v_L}\right]^{\frac{\beta}{\alpha+\beta}}, \tag{31}$$

$$K = Q^{\frac{1}{\alpha+\beta}} \left[\frac{p_K}{p_L}\right]^{\frac{-\alpha}{\alpha+\beta}} \left[\frac{\beta v_K}{\alpha v_L}\right]^{\frac{-\alpha}{\alpha+\beta}}. \tag{32}$$

An implication of the optimization errors, v_{Li} and v_{Ki}, is that the symmetry restriction implied by cost-minimization behavior fails. Specifically, the restriction

$$\frac{\partial L}{\partial p_K} = \frac{\partial K}{\partial p_L} \tag{33}$$

does not hold. Consequently, despite the fact that factor demands honor the feasibility constraint implied by the production function, they do not satisfy all of the restrictions implied by optimizing behavior.

Depending on how optimization errors are introduced, varying degrees of rationality can be imposed on factor demand and consumer demand systems. For example, optimization errors can be introduced in such a way as to yield consumer demands that satisfy the budget constraint and nothing else. This is another way of making Gary Becker's (1962) point that much of the apparent rationality in economic behavior comes from imposing a budget constraint or a technological constraint on what otherwise amounts to irrational behavior.

This discussion of optimization errors has hopefully demonstrated the extremely important and often overlooked point: the addition of disturbances to deterministic behavioral relationships is not innocuous. Depending on how this is done, a well-defined deterministic economic model can be transformed into an incoherent statistical model. For example, if the random disturbances in Equation (26) are allowed to take on values less than zero, for certain realizations of v this system of first-order conditions may not have a solution in x and λ, or may have multiple solutions. Because of these concerns, we recommend that the underlying economic model be formulated with the stochastic structure included, rather than including random shocks into a deterministic model as an afterthought.

4.2.3. Measurement error

Besides these sources of error, structural models also may include measurement errors. Measurement errors occur when the variables the researcher observes are different from those the agents observe. In most cases, it is impossible for researchers to distinguish measurement error from the three other sources of error. As we shall see below, this distinction is nevertheless important, having significant implications not only for estimation and testing, but also for policy evaluations.

Measurement errors also occur in exogenous variables. Unfortunately, these measurement errors often are ignored even though they can be a much greater source of concern. For example, measurement errors in the regressors of a linear regression model will destroy the consistency of OLS. Attempts to handle measurement error in exogenous variables often are frustrated by the fact that there typically is little prior information about the properties of the measurement error. This means that the researcher must predicate any solution on untestable assumptions about the measurement error. As a result, most researchers only acknowledge measurement error in an exogenous variable when they think that the measurement error constitutes a large component of the variation in the exogenous variable.

Measurement error can serve useful purposes in structural econometric modeling. For example, measurement error can make what would otherwise be an incoherent structural model coherent. Consider the case where consumers face nonlinear budget sets. Suppose a consumer must pay $1 per unit for the first 10 units consumed and then $10 per unit for all units beyond the tenth unit consumed. Given the large difference in price between the tenth and eleventh units, we would expect that many consumers would purchase exactly 10 units. In real data, we often do not see dramatic spikes in consumption when marginal prices increase. One way to account for this is to assume that actual consumption is measured with error. This is consistent with the theoretical model's prediction of a probability mass at exactly 10 units, but our not observing a large number of consumers consuming exactly ten units.

Measurement error also is a straightforward way of converting a deterministic economic model into a statistical model. In Example 1, for instance, we introduced measurement errors to justify applying OLS to what otherwise should have been a deterministic relation. However, as we also noted in Example 1, it is usually unrealistic to assume that measurement error is the only source of error. In general, measurement error should be introduced as one of several possible sources of error.

4.3. Steps to estimation

Given a well-defined stochastic model, the next part of our framework is to add any parametric and distributional assumptions necessary to finalize the model. The researcher then is in a position to select an estimation technique and to formulate, where possible, tests of maintained assumptions. We think of this process as having four interrelated selections:

1. selection of functional forms;
2. selection of distributional assumptions;
3. selection of an estimation technique; and
4. selection of specification tests.

There are several criteria a researcher should keep in mind when choosing a functional form. One of the most important is that there is a trade-off between data availability and parametric flexibility. Larger datasets usually allow greater parametric flexibility.

A second criterion is that the functional form should be economically realistic. To take an extreme example, if we are interested in estimating an input elasticity of substitution, then a Cobb–Douglas production function will not work. While this is an extreme case, the structural modeling literature contains nontrivial examples where the functional form almost entirely delivers the desired empirical result.

A third criterion is ease of estimation. If a specific functional form results in a model that is easier to estimate, that should certainly be a factor in its favor. Similarly, if one functional form makes it easier to impose economic restrictions than another, then that too should favor its selection. As an example, it is very easy to impose homogeneity of degree one in input prices on a translog production function. This is not the case for a quadratic cost function. A final criterion is estimation transparency. In some cases, it pays to select a functional form that leads to simpler estimation techniques. This has the advantage of making it easier for other researchers to understand how the researcher arrived at their estimates.

Turning now to the choice of distributional assumptions, a researcher's stochastic specification may or may not involve a complete set of distributional assumptions. To the extent that the researcher is willing to completely specify the distribution of the model errors, the structural model implies a conditional distribution of the observed endogenous variables given the exogenous variables. At this point the researcher can consider using maximum likelihood, or a similar technique (e.g., simulated maximum likelihood or the EM algorithm) to estimate the parameters of interest.

As a specific example, consider an optimizing model of producer behavior. Suppose the economic model specifies a functional form for $\pi(y, x, \epsilon, \beta)$ – a firm's profit function as a function of outputs produced and inputs consumed, y; a vector of input and output prices, x; the vector of firm characteristics observable to the firm but not the researcher, ϵ; and a vector of parameters to be estimated, β. If the firm maximizes profits by choosing y, we have the first-order conditions

$$\frac{\partial \pi(y, x, \epsilon, \beta)}{\partial y} = 0. \tag{34}$$

Assuming that the inverse function $y = h(x, \epsilon, \beta)$ exists and assuming the only source of error, ϵ, has the density, $f(\epsilon, \theta)$, we can apply the change-of-variables formula to compute the density of y from the density of the unobservable ϵ

$$p(y \mid x, \theta, \beta) = f\big(h^{-1}(y, x, \beta), \theta\big)\left|\frac{\partial h^{-1}(y, x, \beta)}{\partial y}\right|. \tag{35}$$

This density can be used to construct the likelihood function for each observation of y.

The final two items on our list include familiar issues in estimation and testing. An advantage of using maximum likelihood in the previous example is that it would be clear to other researchers how the elements of the economic and stochastic models led to the estimation method. There are of course costs to being this complete. One is that maximum likelihood estimators may be difficult to compute. A second is that there is a trade-off between efficiency and robustness. Maximum likelihood techniques may

be inconsistent if not all of the distributional assumptions hold. Generalized method of moments and other estimation techniques may impose fewer restrictions on the distribution of ϵ, but also may yield less efficient estimates. It also is the case that alternatives to maximum likelihood may not allow the estimation of some parameters. This is a corollary to our earlier point about structure. In some instances, the researcher's economic structure exists only because of distributional assumptions. In subsequent sections, we will illustrate how distributional assumptions can identify economic primitives.

Once the researcher obtains estimates of the structural model, it is important to examine, where possible, any restrictions implied by a structural model's economic and stochastic assumptions. In addition, it is useful to examine, where possible, how sensitive estimates are to particular assumptions. Thus, if the researcher has used instrumental variable methods to estimate a model, and there are over-identifying restrictions, then these restrictions should be tested. If a researcher assumes an error term is white noise, then tests for heteroscedastic and/or autocorrelated errors are appropriate. As for the sensitivity of estimates, the researcher can check whether additional variables should be included, or whether other functional form assumptions are too restrictive. Although it is extremely difficult to determine the appropriate nominal size for these specification tests, it is still worthwhile to compute the magnitude of these test statistics to assess the extent to which the structural model estimated is inconsistent with the observed data. Once the structural model is shown not to be "wildly" inconsistent with the observed data, the researcher is ready to use this structural model to answer the sorts of questions discussed in Section 2 and this section.

4.4. Structural model epilogue

An important premise in what follows is that no structural analysis should go forward without a convincing argument that the potential insights of the structural model exceed the costs of restrictive or untestable assumptions. Knowing how to trade off these costs and benefits is critical to knowing whether it makes sense to develop and estimate a structural econometric model. We hope that our framework and our discussion of the IO literature will provide some sense of the "art" involved in building and evaluating structural models.

In what follows, we propose to show how researchers in IO have used structural econometric models. Our purpose is not to provide a complete survey of IO. There already are several excellent literature surveys of areas such as auctions and firm competition. We propose instead to provide a sense of how IO empiricists have gone about combining game-theoretic economic models and statistical models to produce structural econometric models. We also aim to provide a sense of how far IO researchers are in solving important econometric issues posed by game-theoretic models. In our discussions, we hope to convey that structural modeling should be more than high-tech statistics applied to economic data. Indeed, we aim to show through examples how the economic question being answered should motivate the choice of technique (rather than the other way around).

5. Demand and cost function estimation under imperfect competition

In this section, we discuss Porter's (1983) empirical model of competition in an oligopoly market. We begin with Porter's model for several reasons. First, it was one of the first to estimate a game-theoretic model of competition. Second, the model bears a strong resemblance to the classical demand and supply model we discussed in Section 3. Third, we think it is an excellent example of how structural econometric modeling should be undertaken. In the process of reviewing his model, we hope to illustrate how our framework can help identify the essential ingredients of a structural model.

5.1. Using price and quantity data to diagnose collusion

One of the most important research topics in IO is how to measure the extent of competition in an industry. This question is of more than academic interest, as policy makers and the courts often are called upon to assess the extent of intra-industry competition. Additionally, when policymakers or the courts find there is insufficient competition, they must go a step further and propose remedies that will prevent firms from colluding or exercising excessive unilateral market power.

Economists seek to infer the presence or absence of competition from other data, most frequently data on prices and quantities. Sometimes these studies are conducted using firm-level or product-level price and quantity information, and sometimes economists only have industry price and quantity data. The central message of the next several sections is:

> The inferences that IO researchers' draw about competition from price and quantity data rest on what the researchers assume about demand, costs, and the nature of firms' unobservable strategic interactions.

It is therefore essential to evaluate how each of these components affects a researcher's ability to use nonexperimental price and quantity data to identify the extent of competition in an industry.

The demand specification plays a critical role in competition models because its position, shape and sensitivity to competitors' actions affects a firm's ability to mark up price above cost. The IO literature typically draws a distinction between demand models for homogeneous products and differentiated products. In this section we consider homogeneous product models in which firms' products are perfect substitutes and there is a single industry price. In this case, industry demand has the general form

$$Q = h(P, Z, \beta, \nu), \tag{36}$$

where Q is total industry quantity, P is industry price, Z are market demand variables, β are parameters that affect the shape and position of market demand, and ν is a market demand error. This demand function is an economic primitive. By itself it tells us nothing about firm behavior or the extent of competition. Inferences about the extent

of competition, however, are inextricably linked to what the researcher assumes about demand. This is because the demand curve enters into firms' profit-maximizing quantity or price decisions.

To model firms' price or quantity decisions, the researcher must first take a stand on the form of firms' profit functions, specifically the functional forms of market demand and firm-level costs. Once these are specified, the researcher must then introduce assumptions about how firms interact (e.g., Cournot versus Bertrand). Combining these assumptions with the assumption of expected profit-maximizing behavior yields first-order conditions that characterize firms' optimal price or quantity decisions. This "structure" in turn affects the industry "supply" equation that the researcher would use to draw inferences about competition.

In some, but not all, cases it is possible to parameterize the impact of competition on firms' first-order conditions in such a way that they aggregate to an industry price or "supply" equation:

$$P = g(Q, W, \theta, \eta), \tag{37}$$

where W are variables that enter the firms' cost functions, θ are parameters that affect the shape and position of the firms' cost curves and possibly describe their competitive interactions, and η is an error term.

Equations (36) and (37) look like nonlinear versions of the simultaneous linear equations in (3) of Example 3. Both sets of equations describe equilibrium industry prices and quantities. The chief difference is that in an oligopolistic setting, the "supply" equation is not an aggregate marginal cost curve but an aggregation of firm first-order conditions for profit-maximization in which firms mark price up above marginal cost. The extent to which price is above marginal cost depends on firms' competitive interactions. The critical issue is: What about the demand and "supply" equations identifies the extent of competition from observations on prices and quantities?

Porter's study provides a useful vehicle for understanding the assumptions necessary to identify the extent of competition from industry price and quantity data. In particular, his study makes it clear that without imposing specific functional form restrictions on market demand and industry supply, we have no hope of estimating the market demand curve or firm cost curves. This is because the researcher only observes pairs of prices and quantities that solve (36) and (37). Even when the researcher is willing to make distributional assumptions about the joint density of ν and η, without assumptions on the functional form of (36) and (37), the assumption that P and Q are equilibrium magnitudes only implies that there is conditional density of P and Q given Z and W. Consequently, if the researcher is unwilling to make any parametric assumptions for the demand and supply equations, he would, at best, be able to only recover the joint density of P and Q given Z and W using the flexible smoothing techniques described earlier. Only by making parametric assumptions for the supply and demand equations can these two equations be separately identified and estimated from market-clearing prices and quantities. This is precisely the strategy that Porter (1983) and all subsequent

researchers take in estimating the competitiveness of a market from equilibrium price and quantity data.

Rosse (1970) first estimated the extent of unilateral market power possessed by a firm from market-clearing price and quantity, using a sample of monopoly markets. Porter's 1983 study of nineteenth century US railroad cartels is one of the first papers in IO to devise a structural econometric model of a cartelized industry.[3] The economic logic for Porter's empirical model comes from Green and Porter (1984). Green and Porter explore the idea that cartels might use price wars to discipline members who deviate from cartel prices or output quotas. Specifically, Green and Porter develop a dynamic model of a homogeneous product market in which potential cartel members face random shocks to industry demand. By assumption, firms never perfectly observe demand or other firms' output decisions. In this noisy environment, cartel participants have trouble identifying whether lower prices are the result of a breakdown in the cartel or low demand. Green and Porter's work shows that firms can support a cartel by agreeing to a period of competitive pricing of a pre-determined length whenever market prices fall below a trigger price.

In what follows, we use our framework to discuss the components of Porter's model. In particular, we focus on the assumptions that allow Porter to identify competitive pricing regimes. In the process, we hope to illustrate many of our earlier points about structural models. The main lessons we take away from Porter's analysis is that it is impossible to identify the extent of market power exercised by a firm or in an industry from a descriptive data analysis. It is also impossible to determine definitively whether firms are colluding from this sort of data analysis. Inferences about the extent of market power exercised, or the presence and pervasiveness of collusion, rest heavily on untestable economic, functional form and stochastic assumptions. In general, it is not possible to test all these assumptions. The strength of Porter's equilibrium model in which the cartel switches between monopoly and competitive prices is that it is possible to see what is needed to identify monopoly versus competitive regimes.

5.2. The economic model

5.2.1. Environment and primitives

Porter begins, as does most of the structural IO literature, by outlining a static, homogeneous product oligopoly model where the number of firms (entrants) N is exogenously given. All firms know the functional form of market demand and each others' costs. In Porter's homogeneous product model, there is a single, constant elasticity industry demand curve at each period t:

$$\ln Q_t = \alpha + \epsilon \ln P_t + Z_t'\gamma + \nu_t, \tag{38}$$

[3] See Bresnahan (1989) for a detailed survey of early work on estimating market power.

where Q is industry output, P is industry price, Z is a vector of exogenous demand shifters, γ is a conformable vector of unknown coefficients, ϵ is a time-invariant price elasticity of demand, and ν_t is an error term. It appears that Porter uses a constant elasticity demand function because it considerably simplifies subsequent calculations and estimation. Data limitations also limit Z_t to one exogenous variable, a dummy for whether competing shipping routes on the Great Lakes were free of ice. Although he does not discuss the source of the demand error term, it plausible to imagine that it is included to account for demand factors observable to firms but not to Porter.

Each firm has fixed costs of F_i and a constant elasticity variable cost function of the form

$$C_i(q_{it}) = a_i q_{it}^{\delta}, \tag{39}$$

where i indexes firms, t indexes time and q is firm-level output. The motivation for this firm-level cost function appears to be that it delivers an industry "supply" or output curve for a range models of competition.

Porter leaves portions of the economic environment unspecified. Although competing shippers are mentioned, their impact on the railroads is not explicitly modeled. Similarly, although entry by railroads occurs during the sample, the entry decisions are not modeled. (Entry is accounted for by an exogenous shift in the industry supply curve.) Finally, although Porter does not include unobservables in the individual cost functions, as we show below it is possible to rationalize part of the error term that he includes in the industry supply curve as a variable cost component common to all firms that he does not observe.

5.2.2. Behavior and optimization

Porter assumes that each period (one week), firms maximize their per-period profits choosing shipping quantities, q_{it}. Additionally, he assumes each firm forms a conjecture about how other firms will respond to changes in its quantity during that week, θ_{it}. From these behavioral assumptions, Porter derives the standard marginal revenue equals marginal cost quantity-setting first-order conditions for profit maximization by each firm:

$$p_t\left(1 + \frac{\theta_{it}}{\epsilon}\right) = a_i \delta q_{it}^{\delta-1}. \tag{40}$$

Here

$$\theta_{it} = \frac{\partial Q_t}{\partial q_{it}} \frac{q_{it}}{Q_t} = \left(1 + \frac{\partial Q_{-it}}{\partial q_{it}}\right) \frac{q_{it}}{Q_t}$$

and $Q_{-it} = \sum_{k \neq i}^{M} q_{kt}$ is the total amount supplied by all firms besides firm i, and the term $\frac{\partial Q_{-it}}{\partial q_{it}}$ is referred to as firm i's conjectural variation about its competitors' responses to a one unit change in firm i's output level.

Although we discuss conjectural parameters in more detail in the next section, one way to think about the conjectural variation parameter is that it indexes how far price is from marginal cost. If the firm chooses its output assuming it has no influence on market price, then it perceives that any increase in output will be met with an equal and opposite change in the aggregate output of its competitors so that market prices are unchanged. This means $\frac{\partial Q_{-it}}{\partial q_{it}} = -1$, so that θ_{it} equals zero and price equals marginal cost, which implies that the firm assumes it is unable to affect the market price through its quantity-setting actions. For static Cournot–Nash competitors, $\frac{\partial Q_{-it}}{\partial q_{it}} = 0$, which implies that θ_{it} equals firm i's quantity share of the market. For a quantity or price-setting monopoly or cartel, the firm knows that all firms will respond one-for-one with its output change from their current level of output, so that $\frac{\partial Q_{-it}}{\partial q_{it}} = \frac{Q_{-it}}{q_{it}}$, and θ_{it} equals one. This value of θ_{it} implies monopoly pricing on the part of the cartel. Although in principle conjectural variation parameters can continuously range between zero and one, it is unclear what behavioral meaning one would attach to all other values of θ_{it} in this interval besides the three values described above.

While Porter's economic model applies to individual firm decisions, he chooses not to estimate firm-level models. This decision appears to be made because estimating firm-level specifications would add significantly to his computations, particularly if he estimated conjectural variation and cost parameters for each firm. Given the state of computing power at the time he estimated his model, we doubt this would have been computationally feasible. Additionally, such an approach would require him to model new entry during the sample period.

As is common when only industry-level price and quantity data are available, Porter instead aggregates the firm-level first-order conditions to obtain an industry supply equation of the form (37). This approach, while reducing the number of estimating equations, is not without limitations. In aggregating the first-order conditions, it quickly becomes clear that one cannot estimate separate conjectural and cost parameters for each firm and time period. To reduce the dimensionality of the parameters in the industry supply function, Porter assumes that the firm-level values of θ_{it} times the associated market shares are the same (unknown) constant. This assumption has the important computational advantage of reducing the number of conjectural and cost parameters to two. Moreover, it makes it easy to calculate equilibrium prices and quantities in perfectly competitive and monopoly (collusive) markets. It should not be surprising that this simplifying assumption has disadvantages. The two main ones are that the model is now inconsistent with a Cournot market outcome and it is unclear why conjectural parameters should vary inversely with market shares.

Porter obtains his supply equation by weighting each firm's first-order condition in (40) by its quantity,

$$p_t\left(1 + \frac{\theta_t}{\epsilon}\right) = DQ_t^{\delta-1}, \tag{41}$$

where

$$D = \delta\left(\sum_{i=1}^{N} a_i^{\frac{1}{1-\delta}}\right)^{1-\delta}, \tag{42}$$

$$\theta_t = \sum_{i=1}^{N} s_{it}\theta_{it}, \tag{43}$$

and $s_{it} = q_{it}/Q_t$ is the quantity share of firm i in time t. Taking the natural log of this equation yields the aggregate supply function that Porter estimates, apart from the addition of an error term.

At this point, it is useful to summarize Porter's structural model. The main attraction of Porter's assumptions are that they result in a two-equation linear (in the parameters) system that explains equilibrium industry price and quantity data:

$$\begin{aligned} &\ln Q_t - \epsilon \ln p_t = \alpha + Z_t\gamma + \nu_t && \text{Demand Equation,} \\ &-(\delta - 1)\ln(Q_t) + \ln p_t = \lambda + \beta I_t + W_t\phi + \eta_t && \text{Supply Equation,} \end{aligned} \tag{44}$$

where $\lambda = \ln D$, $\beta = -\ln(1 + \theta/\epsilon)$, I_t is an indicator random variable which takes on the value 1 when the industry is in a cooperative regime and 0 when the industry is in a competitive regime, W_t is a set of explanatory variables that capture aggregate supply shifts due to such events as the entry of new firms, and β is an unknown parameter that measures the extent to which price and quantities sold during the collusive regime approach the joint profit-maximizing monopoly solution. For example, if $\beta = -\ln(1 + 1/\epsilon)$, the collusive regime involves joint profit maximization. Lower values of β, however, imply higher output in the collusive regime. Porter argues based on his work with Green, that the true β should be less than the joint profit-maximizing value.

5.2.3. The stochastic model

Porter completes the economic model above with two sets of stochastic assumptions. The first set is fairly standard: he assumes the errors in the demand and industry supply equations are additive, mean zero, homoscedastic normal errors. The source of these errors is left unspecified. One presumes that each error represents demand and cost factors unobservable to modern researchers, but observable to the firms at the time. Porter also assumes the demand and supply errors are independent of the right-hand side exogenous variables. By inspection of the aggregated first-order conditions for profit-maximization in Equation (41), we can see that the supply shock can be rationalized as a common multiplicative supply shock to all firms' variable cost functions. For example, if we redefine a_i in the variable cost function for firm i as $\alpha_{it} = a_i \exp(\eta_t)$, then solving the first-order conditions for each firm and solving for the aggregate supply function, would yield supply functions with the stochastic shock, η_t, given above.

The second stochastic specification Porter adds is less conventional and is motivated by an identification problem. In principle, Porter would like to use data on I_t, which indicates when the cartel was effective, to estimate β (and thereby recover the price–cost markup parameter θ). Unfortunately, he has incomplete historical information on when the cartel was effective. Although he uses some of this information to compare prices and evaluate his model *ex post*, in his main estimations he treats I_t as a random variable that is observable to the firms but not to him. Thus, in effect the error term in the supply equation becomes $\beta I_t + \eta_t$. Absent further information on I_t, it is clear that we have an identification problem – we cannot separately recover the key parameters θ and λ. This problem is akin to having two constant terms in the same regression. To see the problem, notice that the expected value of the error (assuming η_t has mean zero) is $\beta E(I_t)$. This expectation is by assumption nonzero because $E(I_t)$ is the expected value of I_t, which equals the probability that the firms are colluding. Assuming this probability does not change over the sample, which is assumed by Porter's formulation, the nonzero average error is absorbed into the supply equation's constant term, giving $\lambda + E(I_t) = \lambda + \beta\tau$, where τ equals the probability that I_t equals one. The supply disturbance becomes $\beta(I_t - \tau) + \eta_t$. As we can see from the constant term, even if we know the constant β, we cannot separately estimate λ and τ.

To gain another perspective on identification issues in Porter's model, it is useful to compare Porter's model to the linear demand and supply model (3), discussed in the previous section. Porter's demand and supply system has the form

$$
\begin{aligned}
&y_t'\Gamma + x_t'B = E_t', \qquad (45)\\
&\begin{bmatrix} \ln Q_t & \ln p_t \end{bmatrix}\begin{bmatrix} 1 & -(1-\delta) \\ -\epsilon & 1 \end{bmatrix} + \begin{bmatrix} 1 & Z_t' & W_t' \end{bmatrix}\begin{bmatrix} -\alpha & -(\lambda+\beta\tau) \\ -\gamma & 0 \\ 0 & -\phi \end{bmatrix}\\
&\quad = [\nu_t, \beta(I_t - \tau) + \eta_t].
\end{aligned}
$$

Given the parallel, we might be tempted to use the assumptions we applied there, namely that Z_t and W_t are uncorrelated with $\beta(I_t - \tau) + \eta_t$ and ν_t and that the disturbances have a constant covariance matrix. Under these assumptions, we could obtain consistent estimates of the structural parameters, Γ, B and

$$
E\big(E_t E_t'\big) = \Sigma^* = \begin{bmatrix} E(\nu_t^2) & E(\nu_t\eta_t) \\ E(\nu_t\eta_t) & E[(\beta(I_t-\tau)+\eta_t)^2] \end{bmatrix}
$$

in Equation (45) by three-stage least squares.

Notice, however, that in the above formulation, the regime-shift variable I_t only appears in the error term. This suggests that in order to distinguish between Porter's regime-switching model and the classical model, we need to rely on the distributional assumptions Porter imposes on I_t, η_t and ν_t. Absent specific distributional assumptions for I_t and η_t, we have no hope of estimating the probability of regime shifts, τ, or the magnitude of the conduct parameter during these collusive regimes, θ, which is a nonlinear function of β, from the joint distribution of price and quantity data. To identify these parameters, Porter needs to add assumptions. This should not be too surprising given

that he does not observe I_t. His strategy for achieving identification is to parameterize the distribution of the unobservable regimes. Specifically, he assumes that I_t follows an independent and identically distributed (i.i.d.) Bernoulli process and that the ν_t and η_t are i.i.d. jointly normally distributed errors. Further, Porter assumes the demand and supply errors are independent of I_t. These assumptions allow him to identify λ and β separately.

The advantage of Porter's structural framework is that we can explore how these assumptions facilitate identification and estimation. By modeling I_t as an unobservable Bernoulli, Porter has introduced a nonnormality into the distribution of the structural model's errors. To see this, notice that conditional on the regime, the second element of E_t possesses a symmetric normal distribution. Unconditionally, however, the distribution of E_t now is composed of a (centered) Bernoulli and a normal random variable. Consequently, unlike the traditional demand and supply model (3) where we could use standard instrumental variables to recover the relevant structural parameters from conditional mean functions, here we must use more information about the joint distribution of prices and quantities to estimate the model parameters. Put another way, it is the nonnormality of the reduced form errors that determines the extent to which one can identify β empirically. This then raises the delicate question: How comfortable are we with the assumption that η_t and ν_t are normally distributed? Unless there is a compelling economic reason for assuming normality, we have to regard (as Porter does) any inference about regime shifts as potentially hinging critically on this maintained assumption. Fortunately, in Porter's case he does have some regime classification data from Ulen (1978) that agrees with his model's classification of regimes.

At this point it is useful to recall our notion of structure in a simultaneous equations model. As discussed in Section 3, the most that can be identified from descriptive analysis is the conditional density of the endogenous variables, $y_t = (\ln p_t, \ln Q_t')$, given the vector of exogenous variables, $x_t = (1, W_t', Z_t')'$; that is, $f(y_t \mid x_t)$. According to Porter's theoretical model, this observed conditional density is the result of the interaction of industry demand and an industry 'supply' that switches between collusive and noncooperative regimes. However, no amount of data will allow the researcher to distinguish between this regime-switching structural model and a conventional linear simultaneous equations model with no regime switching.

To derive the likelihood function for the case of a single regime linear simultaneous equation model, consider the error vector in Equation (45). The first error is by assumption a mean-zero normal random variable and the second is the sum of a centered Bernoulli random variable, $I_t - \tau$ and a mean zero normal random variable. Applying the law of total probability formula yields the following density for E_t

$$g(E_t) = \tau \frac{1}{2\pi} |\Sigma|^{-1/2} \exp\left(-\frac{F_{1t}' \Sigma^{-1} F_{1t}}{2}\right) + (1-\tau) \frac{1}{2\pi} |\Sigma|^{-1/2} \exp\left(-\frac{F_{2t}' \Sigma^{-1} F_{2t}}{2}\right),$$

where

$$F_{1t} = \begin{bmatrix} E_{1t} \\ E_{2t} - \beta(1-\tau) \end{bmatrix} \quad \text{and} \quad F_{2t} = \begin{bmatrix} E_{1t} \\ E_{2t} + \beta\tau \end{bmatrix}.$$

Both models give rise to the same conditional density $f(y_t \mid x_t)$, but have very different economic implications. The first model implies random switches from competitive to collusive pricing regimes; the other implies a single-pricing regime but a nonnormal distribution for E_t. Consequently, any test for regime shifts must be conditional on the assumed supply and demand functions, and more importantly, the assumed distributions for I_t and η_t. Because these distributional assumptions are untestable, as this example illustrates, we believe that any test for stochastic regime shifts, should be interpreted with caution.

One might view this result as a criticism of structural modeling. To do so would miss our earlier points about the strengths of a structural model. In particular, a key strength of a structural model is that it permits other researchers to ask how the modeler's assumptions may affect results. This example also illustrates our earlier meta-theorem that: absent assumptions about the economic model generating the observed data, the researcher can only describe the properties of the joint distribution of x_t and y_t.

To understand all of the implications of this point, we re-write Porter's regime switching model as

$$y_t'\Gamma = x_t'D + I_t\Delta + U_t', \tag{46}$$

where

$$\Gamma = \begin{bmatrix} 1 & -(1-\delta) \\ -\epsilon & 1 \end{bmatrix}, \quad \Delta = [\,0 \quad \beta\,], \quad D = \begin{bmatrix} \alpha & \lambda \\ \gamma & 0 \\ 0 & \phi \end{bmatrix}, \quad U_t = \begin{bmatrix} \nu_t \\ \eta_t \end{bmatrix},$$

$$\text{and} \quad U_t \sim N(0, \Sigma). \tag{47}$$

In terms of this notation, the conditional density of y_t given x_t and I_t is:

$$h(y_t \mid I_t, x_t) = \frac{1}{2\pi}|\Sigma|^{-1/2} \times \exp\left(-\frac{(y_t'\Gamma - x_t'D - I_t\Delta)\Sigma^{-1}(y_t'\Gamma - x_t'D - I_t\Delta)'}{2}\right).$$

Using the assumption that I_t is an i.i.d. Bernoulli random variable distributed independent of U_t and x_t yields the following conditional density of y_t given x_t:

$$f(y_t \mid x_t) = \tau h(y_t \mid I_t = 1, x_t) + (1-\tau)h(y_t \mid I_t = 0, x_t).$$

As has been emphasized above and in Section 3, all that can be estimated from a statistical analysis of observations on x_t and y_t is the true joint density of $f^{\text{true}}(y_t, x_t)$, from which one can derive the conditional density of y_t given x_t. The fact that $f^{\text{true}}(y_t \mid x_t)$, the true conditional density, can be factored into the product of two conditional normal densities times the probability of the associated value of I_t is due solely to the

functional form and distributional assumptions underlying Porter's stochastic economic model.

Without imposing this economic structure on $f(y_t \mid x_t)$, the researcher would be unable to estimate underlying economic primitives such as the price elasticity of demand, the price elasticity of supply, the probability of a collusive versus a competitive regime, and the magnitude of the difference in prices between the collusive and competitive regimes. Even the best descriptive analysis would yield little useful economic information if the true data-generation process was Porter's structural model. Suppose that one had sufficient data to obtain a precise estimate of $f^{\text{true}}(y_t, x_t)$ using the techniques in Silverman (1986). From this estimate, the researcher could compute an estimate of $E(y_t \mid x_t)$ or the conditional density of y_t given x_t. However, suppose the researcher computed $\frac{\partial E(y_t|x_t)}{\partial x_{it}}$ for the ith element of x_t. If Porter's model were correct, this expectation would equal

$$\tau \frac{\partial E(y_t \mid I_t = 1, x_t)}{\partial x_{it}} + (1-\tau)\frac{\partial E(y_t \mid I_t = 0, x_t)}{\partial x_{it}},$$

so that any partial derivative of the conditional mean is an unknown weighted sum of partial derivatives of the conditional means under the competitive and collusive regimes. The researcher would therefore have a difficult time examining the validity of comparative statics predictions concerning signs of these partial derivatives under competition versus collusion, unless the sign predictions were the same under both regimes. Inferring magnitudes of the competitive or collusive comparative static effects, would be impossible without additional information.

This last observation raises an important point about the success we would have in trying to enrich the economic model of regime shifts. Imagine, as some have, that there are more than two regimes. We might attempt to model this possibility by assuming that I_t has multiple points of support. This seemingly more reasonable model imposes greater demands on the data, as now the extent to which these additional supply regimes are "identified" is determined by a more complicated nonnormal structure of the reduced form errors.

One final point about the estimation of β is that care must be exercised in drawing inferences about the presence of multiple regimes. Under the null hypothesis that there are no regime shifts, standard likelihood ratio tests are invalid. The problem that arises is that under the null of no regime shifts, τ, the probability of the collusive regime, is equal to zero and β is no longer identified. Technically this causes problems because the information matrix is singular when $\tau = 0$. It is unclear then what meaning we can attach to standard tests of the hypothesis that there are distinct regimes.

5.3. Summary

Our analysis of Porter's model leads us to conclude that demand and supply models for oligopolistic industries pose special identification and applied econometric problems. More importantly, the parameters describing competitive conjectures or the degree of

competition are not necessarily identified with commonly available data. In general, the researcher will have to have within-sample variation in demand or cost parameters, or make specific distributional assumptions and apply specific estimation techniques, to identify how competitive conduct affects industry supply behavior. As we shall see, this identification problem is all too common in industrial organization models of firm and industry behavior.

The strength of Porter's model is that it both illustrates potential identification and estimation problems posed by the standard theory and commonly available industry data. It also provides a strategy for recovering information about competitive regimes from limited information about the prevailing competitive regime. Although one could consider alternative strategies for identifying the competitive regimes, Porter compares his estimates of the probability of collusion to information from Ulen (1978) on when the cartel was actually effective. This is an example of how other evidence can be brought to bear to check whether the results of the structural model make sense. Porter finds a remarkable amount of agreement between the two measures. His model also provides an economically plausible explanation for the enormous variation in grain prices over his sample period.

6. Market power models more generally

Porter's model is an example of an IO model that uses data on market-clearing prices and outputs to draw inferences about the extent of market competition. Because these are among the most widely used empirical models in industrial organization, it is worth going beyond Porter's model to consider what other studies have done to identify market power. There are an enormous number of market power studies, many more than we can do justice to here. Bresnahan (1989) surveys the early papers in this area. Our focus is on illustrating the critical modeling issues that arise in the identification and estimation of these models.

Most empirical researchers in IO define a competitive market outcome as one where price equals the marginal cost of the highest cost unit supplied to the market. If the market price is above this marginal cost, then firms are said to exercise "market power". While some studies are content simply to estimate price–cost margins, many go further and attempt to infer what types of firm behavior ("conduct") are associated with prices that exceed marginal costs. A first observation we make below is: absent a structural model, one cannot infer the extent of competition from the joint distribution of market-clearing prices and quantities. Put another way, one needs an economic model to estimate marginal costs (and hence price–cost margins) from the joint distribution of market-clearing prices and quantities. This structural model will involve functional form assumptions and often distributional assumptions that cannot be tested independently of hypotheses about competition.

While this observation may seem obvious from our discussion of Porter's model, there are plenty of examples in the literature where researchers draw unconditional in-

ferences about the extent of competition. That is, they draw inferences about price–cost margins without acknowledging that their inferences may depend critically on their economic and functional form assumptions.

A second observation below is: while one can estimate price–cost margins using a structural model, it is problematic to link these margins to more than a few specific models of firm behavior. In particular, many studies estimate a continuous-valued parameter that they claim represents firm "conjectures" about how competitors will react in equilibrium. Currently there is no satisfactory economic interpretation of this parameter as a measure of firm behavior – save for firms in perfectly competitive, monopoly, Cournot–Nash and a few other special markets. We therefore see little or no value to drawing economic inferences about firm conduct from conjectural variation parameter estimates.

In what follows we discuss these two observations in more detail. We first discuss how the literature identifies and interprets market power within the confines of static, homogenous goods models where firms choose quantities. We then discuss at a broader level what market power models can tell us in differentiated product markets.

6.1. Estimating price–cost margins

Since the late 1970s, many papers in IO have used firm and industry price and quantity data to describe competition in homogeneous product markets. The typical paper begins, as Porter did, by specifying a demand function and writing down the first-order condition:

$$P + \theta_i q_i \frac{\partial P}{\partial Q} = \mathrm{MC}_i(q_i). \tag{48}$$

The goal of these papers is to estimate the 'conduct' parameter θ_i. Most authors assert that this parameter measures firm "conjectures" about competitor behavior. As such, it would seem to be a structural parameter that comes from an economic theory. Is this the case?

Isolating θ_i in Equation (48), and letting α_i denote firm i's output share and ϵ the elasticity of demand, we obtain

$$\theta_i = \frac{P - \mathrm{MC}_i(q_i)}{-q_i \frac{\partial P}{\partial Q}} = \frac{P - \mathrm{MC}_i(q_i)}{P} \frac{1}{\alpha_i \epsilon}. \tag{49}$$

From this equation, we see that θ_i provides essentially the same *descriptive* information as Lerner's (1934) index. That is, it provides an idea of how far a firm's price is from its marginal cost. To the extent that price is above marginal cost (i.e., the Lerner index is positive), IO economists claim that the firm has 'market power'.

Equation (49) is useful because it identifies two critical structural quantities that a researcher must have to estimate θ_i. These are the price elasticity of demand and marginal cost. Following Porter, a researcher could in principle separately estimate the price elasticity of demand from price and quantity data. In developing such an estimate, the

researcher would of course have to worry that the demand function's form may critically impact the estimated elasticity. The marginal cost term in Equation (49) poses a more difficult estimation problem. Equation (49) tells us that with just price and quantity data, we cannot separate the estimation of marginal cost from the estimation of θ_i. Even if we have observations on total or even variable cost associated with this level of output, we are unable to separate them without making specific functional form assumptions for demand and marginal cost. Put another way, the identification of θ_i hinges on how we choose to estimate marginal cost and the aggregate demand curve. Changing the marginal cost and demand specification will change our estimate of θ_i. Unless one knows the functional form of demand and costs, it is impossible to determine the value of θ_i.

Despite the many untestable functional form assumptions necessary to infer marginal costs from price and quantity data, many studies go further and use Equation (48) to estimate θ_i and interpret it as a measure of firm behavior. To understand where this behavioral interpretation comes from, we return to the economic rationale underlying Equation (48). In Equation (48), θ_i is a placeholder for the derivative:

$$\theta_i = \frac{dQ}{dq_i}. \tag{50}$$

According to this definition, θ_i is not a statement about how far prices are from marginal costs, but rather a "variational" concept associated with firm behavior. Specifically, Equation (48) sometimes is interpreted as saying: the firm "conjectures" industry output will increase by θ_i should it increase its output by one unit. The problem with this interpretation is that there are only a few values of θ_i where economists have a good explanation for how firms arrived at such a conjecture. This leads to our second observation above. We know of no satisfactory static model that allows for arbitrary values of θ_i. Empirical models that treat θ_i as a continuous value to be estimated thus are on shaky economic ground, particularly because these estimates of θ_i are predicated on a specific functional form for marginal costs and demand.

To emphasize the danger inherent in associating residually determined θ_i with behavior, imagine observing two firms producing different quantities who otherwise appear identical. The conjectural variation approach would explain the difference by saying firms simply "expect" or "conjecture" that their competitors will react differently to a change in output. Yet there is no supporting story for how otherwise firms arrived at these different conjectures. On the other hand, even though the firms appear identical, one might wonder whether their marginal costs are identical. It seems plausible to us that unobservable differences in marginal costs, rather than behavior, could explain the difference in output. Absent a richer model of behavior that explains where conjectures come from, it is anyone's guess.

To summarize our discussion so far, we have provided two possible interpretations of θ_i. There are, however, a few instances in which θ_i sensibly corresponds to a specific market equilibrium. A leading case is price-taking competition, where $\theta_i = 0$ and price equals marginal cost. Cournot ($\theta_i = 1$), Stackleberg and monopoly are three other

well-known cases. While there has been some debate in the theoretical literature about whether these models are internally "consistent" static behavioral models [e.g., Lindh (1992)], each of these models lends itself to a natural interpretation of what θ_i means as a conjecture about competitor behavior. Thus, it seems to us sensible to imagine imposing these conjectures in the first-order condition (48) and using them to estimate the parameters of demand and cost functions. One can use nonnested tests, as in Bresnahan (1987), to determine which of these different models of behavior are most consistent with the joint density of the data.

Having said this, we realize that some might argue that one loses little by treating θ_i as a continuous parameter to be estimated. After estimating it, the argument goes, one can still compare it to the benchmark values. For example, suppose one precisely estimated $\theta_i = 1.7$, and could reject perfect competition and Cournot. One might think it reasonable to conclude the market is "less competitive than Cournot". But does this make much sense? According to the conjectural variations story, and Equation (48), an estimate of 1.7 implies that firm i believes that if it increases output by one unit, industry output will increase by 1.7 units. What type of behavior or expectations leads to firm i maximizing its profits by maintaining $\theta_i = 1.7$? Why does this value not simply reflect the extent of misspecification of the demand and cost functions in a Cournot model? The problem here is that the theory underlying firm i's behavior (and those of its competitors' behavior) is static. There is no obvious explanation for why firm i has this behavior. Moreover, as we show in the next section, in order to identify an estimate of θ_i, a researcher must select a parametric aggregate demand curve and rule out several types of functional forms for aggregate demand. Otherwise it is impossible to identify θ_i from market-clearing price and quantity data.

If there is an answer to the question of where a firm's conjectures comes from, it must come from a dynamic model of "conjectures" formation. Riordan (1985) provides one such model. Given the subtleties involved with reasoning through how today's competitive interactions might affect future beliefs, it seems unlikely dynamic models will produce simple parameterizations of conjectures or easily estimated first-order conditions. Moreover, the literature on repeated games has shown that when modeling current behavior, one has to recognize that threats or promises about future behavior can influence current behavior. This observation points to a distinction between what firms do in equilibrium (how they appear to "behave") and what they conjecture their competitors' would do in response to a change in each firm's output.[4] This also is a distinction that Stigler (1964) used to criticize static conjectural variation models.

To understand how this distinction affects empirical modelers, consider a cartel composed of N symmetric firms, each of whom charges the monopoly price. In this case, one would estimate θ_i equal to the number of firms. If we gave this estimate a behavioral interpretation, we would report that in this industry, firms conjecture or expect other firms to change their outputs one-for-one. Yet this may not be the case at all, as

[4] Corts (1999) makes a similar argument.

some recent theories have emphasized. The firms may be charging the monopoly price because they expect that if they defect from the monopoly price by producing a little more, each of their competitors may punish them by producing much more.

This distinction between the "beliefs" that economic agents hold and what they ultimately may do in equilibrium is critical for exactly the reasons we outlined in our introductory framework. If one wants to describe where price is in relation to a firm's marginal cost, then θ_i provides a descriptive measure of that, but not a statement about behavior. If, however, one wants to use the estimated parameters to predict what would happen if the firms' economic environment changes, then one either must have a theory in which beliefs and equilibrium behavior coincide, or one must ask which of a small set of values of θ_i, corresponding to perfect competition, monopoly, Cournot and the like, best explains the data.

6.2. *Identifying and interpreting price–cost margins*

In the previous subsection we emphasized that while one could relate θ to price–cost margins, one could not separately estimate θ and marginal costs from price and quantity data alone. Despite occasional claims to the contrary, assumptions about the functional form of marginal costs are likely to affect estimates of θ and vice versa. This section illustrates how assumptions about the structure of demand and marginal costs impact the estimation of the descriptive parameter θ. (Throughout this subsection, we think of θ as providing descriptive information about price–cost margins.)

The IO literature has adopted different approaches to estimating price–cost margins depending upon whether or not they have individual firm or industry price and quantity data. When only industry-level data are available, researchers typically use the equation

$$P + \theta Q \frac{\partial P}{\partial Q} = \mathrm{MC}(Q) \tag{51}$$

to estimate a single industry θ. James Rosse's (1970) paper is the first to estimate the degree of market power (the price–cost markup), or equivalently a firm's marginal cost curve. He used observations on market-clearing prices and quantities from a cross-section of US monopoly newspaper markets. Rosse's procedure uses this first-order condition with θ set equal to 1, along with an assumed parametric aggregate demand curve to estimate the marginal cost curve. This procedure works for the following reason. Once a parametric functional form for demand is selected, this can be used to compute $\frac{\partial P}{\partial Q}$ for each observation in the sample. Setting the value of θ for each observation to 1 guarantees that we have the information necessary to compute the left-hand side of Equation (51) for each observation. This provides an implied value of marginal cost for every output level in the sample. Combining this data with a parametric specification for the firm's marginal cost function, we can estimate marginal cost parameters.

To extend Equation (51) to an oligopoly market requires further assumptions. This equation would appear to mimic a single firm's first-order condition, and thus we might

think of it as linked to the price–cost margins of a "representative" firm. But this is not generally true. Starting as Porter did from the individual firm profit maximization conditions, we can sum Equation (48) across firms to obtain the relation

$$P + \frac{\partial P}{\partial Q} \sum_{i=1}^{N} \frac{\theta_i q_i}{N} = \sum_{i=1}^{N} \frac{\mathrm{MC}(q_i)}{N}, \tag{52}$$

which we can rewrite as

$$P + \theta \frac{\partial P}{\partial Q} Q = \overline{\mathrm{MC}(q_i)}. \tag{53}$$

Here, $\theta = \frac{1}{N} \sum_{i=1}^{N} \frac{\theta_i q_i}{Q}$ is an average of firm market shares times the individual firm θ_i parameters, and $\overline{\mathrm{MC}(q_i)}$ is the average of the N firms' marginal costs. While this equation "looks" like the industry aggregate equation (51) used in many studies, it is not the same without further assumptions. Note, for example, that if θ_i varies across firms, then changes in firms' market shares will generally change θ. Thus, if one is analyzing time series data on prices and output, it may make little sense to treat θ in Equation (51) as a constant. An exception is when one assumes all firms have the same θ_i. But in this case, one must have the same number of firms in the industry for θ to remain constant through time.

The assumption that all firms have the same θ_i amounts to assuming that at the same production level, all firms in the industry would have similarly sloped firm-level demand curves and the same marginal revenues. This is a nontrivial restriction which would require justification on a case-by-case basis. A number of studies, beginning with Gollop and Roberts (1979), Applebaum (1982) and Spiller and Favaro (1984), have argued that one should relax this restriction by making θ a function of different variables, including output. To date, however, there is very little economic theory to guide structural models of how θ_i varies across firms. The most widely adopted specifications are ad hoc, with θ depending on firm output, market share or a firm's size rank.

Another consequence of assuming all firms have the same θ is that differences in firms' outputs now are a function solely of differences in marginal costs. In some instances, this leads to a monotonic relationship between the efficiency of a firm and its observed production. For example, if we assume marginal costs are increasing in output, then there is an inverse relationship between output and marginal costs. Thus, the firm with the largest output has the lowest marginal cost, the firm with the second largest output the second lowest marginal cost, and so on. While this relationship may be entirely reasonable for many industries, it may not be for all.

Turning now to the right-hand side of Equation (51), we see that the notation $\mathrm{MC}(Q)$ gives the impression that only industry output enters the industry supply relation. Put another way, a reallocation of output from one firm in the industry to another will not change the right-hand side of the industry supply relation (51). This obviously cannot generally be true. Equation (53) shows why this is so. To explore this point further, it is

useful to assume that firms have linear marginal costs of the form

$$\mathrm{MC}(q_i) = c_{0i} + c_{1i} q_i. \tag{54}$$

In this case, we can rewrite Equation (53) as

$$P + \tilde{\theta} Q \frac{\partial P}{\partial Q} = \bar{c_o} + \tilde{c_1} Q + \psi, \tag{55}$$

where

$$\tilde{\theta} = \frac{\sum_{i=1}^{N} \frac{\theta_i}{N}}{N}, \tag{56}$$

$$\bar{c_0} = \frac{1}{N} \sum_{i=1}^{N} c_{0i}, \qquad \tilde{c_1} = \frac{1}{N^2} \sum_{i=1}^{N} c_{1i}, \tag{57}$$

$$\psi = \mathrm{Cov}(c_{1i}, q_i) - \mathrm{Cov}(\theta_i, q_i) \frac{\partial P}{\partial Q} \tag{58}$$

and $\mathrm{Cov}(x, y)$ equals the covariance (calculated over firms in the industry) between x and y. If the ψ term is zero, then Equations (53) and (51) are indistinguishable. This happens for example when firms have similarly sloped marginal cost functions and the same θ. In general, however, we can think of Equation (51) as having an error term that includes ψ. To the extent that ψ is nonzero and varies systematically in the researcher's sample, the researcher will obtain biased estimates of the demand, cost and θ parameters by ignoring ψ.

We now turn to considering whether and how functional form assumptions might affect inferences about θ from industry price and quantity data. Both Bresnahan (1982) and Lau (1982) consider the issue of identification in detail using the aggregate equation (51). Because their results apply to a special aggregation of individual firm first-order conditions, it is useful to revisit their discussion in the context of the individual firm marginal revenue equal to marginal cost conditions. To facilitate this discussion, let each firm face the demand function $Q = D(P, Y, \alpha)$, where α is a vector of demand parameters and Y is a set of exogenous variables that shift demand but not cost. Suppose also that each firm has the marginal cost function $\mathrm{MC}_i = c_0 + c_1 q_i + c_2 w_i$, where w_i is an exogenous cost shifter. If a researcher had time series data on market prices, firm i's output, Y and w_i over time, the researcher could estimate firm i's market power parameter θ_i using the two equation system

$$\begin{aligned} Q &= D(P, Y, \alpha), \\ P &= c_0 + \left(c_1 + \frac{\partial D^{-1}}{\partial Q} \theta_i \right) q_i + c_2 w_i \end{aligned} \tag{59}$$

once some assumption had been made about unobservables. The second equation shows that by assuming marginal costs are linear in output, we have potentially destroyed the identification of θ_i. Consider, for example, what happens when demand has the form

$Q = \alpha_0 + \alpha_1 P + \alpha_2 Y$. In this case, firm i's supply relation is

$$P = c_0 + \left(c_1 + \frac{\theta_i}{\alpha_1}\right) q_i + c_2 w_i. \tag{60}$$

Hence, even though we can obtain a consistent estimate of the demand parameter α_1 from the demand equation, we cannot separate c_1 from a constant θ_i. Of course, if we are willing to restrict θ, we can identify the marginal cost parameters and price–cost margins.

It is tempting to identify θ_i in this case by assuming that marginal costs are constant, i.e., $c_1 = 0$. Unfortunately, researchers rarely have independent information that would support this assumption. Alternatively, following Bresnahan (1982), one could identify θ_i by allowing the slope of market demand to vary over time in an observable way. For instance, one might interact price with income (Y) in the demand equation

$$Q = \alpha_0 + \alpha_1 P + \alpha_2 Y P$$

to obtain the supply equation

$$P = c_0 + \left(c_1 + \frac{\theta_i}{\alpha_1 + \alpha_2 Y}\right) q_i + c_2 w_i. \tag{61}$$

Although θ_i is formally identified in this specification, its identification in practice depends heavily on having variables, such as income, that interact or otherwise cannot be separated from price [e.g., Lau (1982)]. In other words, the value of θ is identified off of a functional form assumption for aggregate demand.

Yet another approach to identifying θ_i that has not been fully explored is to add information from other firms' supply relations. In the language of econometrics, it may be possible to obtain identification by imposing cross-equation restrictions between the pricing equations. Returning to the specification in Equation (53), if we added a supply curve for a second firm j, we still would not be able to identify θ_i or θ_j. We would, however, be able to identify the difference if we assumed that both firms' marginal cost functions had the same slope. Alternatively, we could identify the difference in the slopes of the firms' marginal cost functions if in a panel data setting (where T goes to infinity) we assume that all firms have the same constant θ.

Our discussion so far has suggested that θ is identified by the functional form assumptions one makes about market demand and firms' costs. This dependence seems to not always be appreciated in the literature, where cost and demand functions are sometimes written down without much discussion of how their structure might affect estimates of θ. A useful example of how the functional form of demand affects the identification of θ is provided by the inverse demand function:

$$P = \alpha - \beta Q^{1/\gamma}. \tag{62}$$

This inverse demand function leads to the direct estimator (by applying Equation (49) above)

$$\theta_1 = -\gamma \frac{P - c}{\alpha - P}, \tag{63}$$

which illustrates how the demand parameters affect the direct estimate. This inverse demand function also yields a transformed Equation (51)

$$P_t = \frac{\gamma c_t}{\gamma + \theta} + \frac{\alpha\theta}{\gamma + \theta}, \tag{64}$$

where the subscript t denotes variables that are naturally thought of as time varying. Critical to most applications is what one assumes about marginal costs. In the simplest case, one can think of firms as having constant, but time-varying marginal costs c_t which depend linearly on some time-varying exogenous covariates, i.e.,

$$c_t = c_0 + W_t\omega,$$

where ω is a vector of parameters. Substitution of this relationship into (64) gives the equation

$$P_t = \frac{\alpha\theta}{\gamma + \theta} + \frac{\gamma c_0}{\gamma + \theta} + \frac{\gamma}{\gamma + \theta} W_t\omega = \beta_0 + \beta_1 W_t.$$

This equation makes it clear that absent further assumptions, we cannot identify θ from estimates of β_0 and β_1 alone. One way around this problem is to recognize from Equation (53) that θ depends on market shares and the number of firms, both of which are potentially time varying. This, however, is not the usual approach. Instead, most studies follow the advice of Bresnahan and Lau and identify θ by assuming that the demand parameters α and/or γ contain a demand covariate. For example, if we assume that the inverse demand intercept equals

$$\alpha_t = \alpha_0 + D_t\alpha_1,$$

then Equation (64) becomes

$$P_t = \frac{\alpha_0\theta}{\gamma + \theta} + \frac{\gamma c_0}{\gamma + \theta} + \frac{\alpha_1\theta}{\gamma + \theta} D_t + \frac{\gamma}{\gamma + \theta} W_t\omega.$$

This equation and the demand equation now exactly identify θ. But note that the estimate of θ depends critically on the effect of D on demand and on the curvature of demand. If we had started out, as many studies do, by assuming linear demand then we could not estimate θ. This is yet another example of how economic structure is solely identified by a functional form or distributional assumption.

6.3. Summary

In this section we have discussed how IO researchers use price and quantity data to estimate price–cost margins. We also have questioned the value of static conjectural variation parameters. Apart from these observations, we have tried to underscore one of the key observations of our framework, which is that functional form assumptions play a critical role in inferences about marginal economic "effects" and the appropriate model of competition.

7. Models of differentiated product competition

The previous two sections discussed how IO economists have used price and quantity data to draw inferences about the behavior of oligopolists selling homogeneous products. These empirical models parallel textbook demand and supply models. The chief difference is in an oligopoly model, the supply equation is replaced by a price equation derived from first-order conditions that describe how oligopolists maximize profits. Because IO economists do not observe the marginal costs that enter these first-order conditions, they are forced to estimate them along with other structural parameters. It should not be too surprising that a researcher's stochastic and functional form assumptions have a critical impact on the resulting estimates, as the researcher is simultaneously trying to draw inferences about the nature of demand, costs and competition from just data on prices and quantities.

This section examines how IO economists have used price and quantity information on differentiated products to draw inferences about demand, costs and competition. We first discuss complexities that arise in neoclassical extensions of homogeneous product models. We then turn to more recent differentiated product discrete choice models.

7.1. Neoclassical demand models

In the late 1980s and 1990s, empirical IO economists began to focus on modeling competition in differentiated product markets such as cars, computers and breakfast cereals. Bresnahan (1981, 1987) are two early examples of this work. These models also use price and quantity data to draw inferences about oligopolists' strategic interactions and price–cost markups. The main difference between these models and homogeneous product models is that the researcher specifies separate "demand" and "supply" equations for each product. Thus, instead of working with two-equation, market-level systems such as (36) and (37), the researcher specifies a J-product demand system:

$$\begin{aligned} Q_1^d &= h_1(P_1, P_2, \ldots, P_J, Z_1, \beta_1, \nu_1), \\ &\vdots \\ Q_J^d &= h_J(P_1, P_2, \ldots, P_J, Z_J, \beta_J, \nu_J) \end{aligned} \tag{65}$$

and a J-equation system of first-order profit maximization conditions:

$$\begin{aligned} P_1 &= g_1\big(Q_1^s, Q_2^s, \ldots, Q_J^s, W_1; \theta_1, \eta_1\big), \\ &\vdots \\ P_J &= g_J\big(Q_1^s, Q_2^s, \ldots, Q_J^s, W_J; \theta_J, \eta_J\big). \end{aligned} \tag{66}$$

Although these systems look much more complicated than the simultaneous equations in the homogenous product case, they pose the same basic modeling issue: unless the researcher is willing to make specific functional form assumptions for firms' demands

and costs, the researcher will be unable to draw inferences about equilibrium firm-level markups. This issue arises again because, absent economic assumptions about the structure of demand and costs, the most the researcher can do is use flexible data-smoothing techniques to recover the conditional joint density of the J prices and J quantities given the demand and cost variables W and Z. Only by making functional form assumptions for demand and costs, and assumptions about the forms of strategic interactions, can the researcher recover information about demand and cost primitives. This means that we still need specific functional form assumptions to use price and quantity data to draw inferences about equilibrium markups.

The main new issue posed by differentiated products is one of scale. Now, the researcher has to specify a set of demand functions – potentially involving dozens or hundreds of products. Left unrestricted, the number of parameters in these demand systems can easily exceed the number of observations in conventional market-level price and quantity datasets. This problem has led IO researchers to focus on how best to formulate parsimonious, yet flexible, demand systems.

To appreciate the practical issues involved, consider the challenge IO economists or antitrust authorities face in trying to assess the competitiveness of the US ready-to-eat breakfast cereal industry. Absent cost data, inferences about manufacturer and retailer price–cost margins have to be drawn from retail prices and sales. As there are over 50 major brands of cereals, a simple model would have at least 100 equations – 50 demand and 50 "supply" equations. Each equation conceivably could contain dozens of parameters. For instance, paralleling Porter's homogeneous product specification, we could assume a log-linear demand system:

$$\begin{aligned}
\ln Q_1 &= \beta_{10} + \beta_{11}\ln y + \beta_{12}\ln P_1 + \beta_{13}\ln P_2 + \cdots + \beta_{1,51}\ln P_{50} + Z_1\gamma_1 + \nu_1,\\
\ln Q_2 &= \beta_{20} + \beta_{21}\ln y + \beta_{22}\ln P_1 + \beta_{23}\ln P_2 + \cdots + \beta_{2,51}\ln P_{50} + Z_2\gamma_2 + \nu_2,\\
&\vdots\\
\ln Q_{50} &= \beta_{50,0} + \beta_{50,1}\ln y + \beta_{50,2}\ln P_1 + \beta_{50,3}\ln P_2 + \cdots\\
&\quad + \beta_{50,50}\ln P_{50} + Z_{50}\gamma_{50} + \nu_{50}.
\end{aligned} \tag{67}$$

This system has over 2600 parameters![5] Such unrestricted parameterizations easily exceed the number of observations obtainable from public sources.

Even when the researcher has large amounts of data, Equations (65) and (66) pose significant computational challenges. For instance, to use maximum likelihood, the researcher would have to work with the Jacobian of 100 demand and markup equations. Nonlinearities in the system also raise the concern that the system may not have a unique solution or a real-valued solution for all error and parameter values. Although these complications can sometimes be dealt with in estimation, they may still reappear when the researcher performs counterfactual calculations. For instance, there may be

[5] Recall that aggregate demand need not be symmetric.

no nonnegative prices or single set of prices that solve (65) and (66) for a particular counterfactual.

These econometric issues have prompted IO researchers to look for ways to simplify traditional neoclassical demand models. Many early simplifications relied on ad hoc parameter restrictions or the aggregation of products.[6] For example, to estimate (67) a researcher might constrain a product's cross-price elasticities to all be the same for all products.[7] Simplifications such as this, while computationally convenient, can unduly constrain estimates of price–cost markups.

Multi-level demand specifications provide a somewhat more flexible demand function parameterization.[8] In a multi-level demand specification, the researcher separates the demand estimation problem into several stages or levels. At the highest level, consumers are viewed as choosing how much of their budget they wish to allocate to a type of product (e.g., cereal). At the next stage, the consumer decides how much of their budget they will divide among different categories of the product (e.g., categories of cereal such as kids', adult and natural cereals). At the final stage, the consumer allocates the budget for a category among the products in that category (e.g., within kids' cereals, spending on Trix, Count Chocula, etc.).

Although multi-stage models also restrict some cross-price elasticities, they permit flexible cross-price elasticities for products within a particular product category. For example, a researcher can estimate a flexible neoclassical demand system describing the demands for kids' cereal products. Changes in the prices of products in other categories (e.g., adult cereals) will still affect the demands for kids' cereals, but only indirectly through their effect on overall kids' cereal spending. Exactly how these restrictions affect estimates of markups is as yet unclear.[9]

Other recent work in the neoclassical demand system tradition has explored reducing the number of demand parameters by constraining cross-price effects or making them depend on estimable functions of covariates.[10] Pinkse, Slade and Brett (2002), for example, constrain the coefficients entering firms' price elasticities to be functions of a small set of product attributes. While this strategy facilitates estimation and allows flexibility in own and cross-price effects, it has the disadvantage of being ad hoc. For instance, it is not clear where the list of attributes comes from or how the functional

[6] Bresnahan's (1989) Section 4 reviews early efforts. Deaton and Muellbauer (1980) provide a survey of neoclassical demand models.

[7] One utility-theoretic framework that produces this restriction is to assume that there is a representative agent with the constant elasticity of substitution utility function used in Dixit and Stiglitz (1977).

[8] See, for example, Hausman, Leonard and Zona (1994).

[9] Theoretical work, beginning with Gorman (1959), has explored the restrictions that multi-stage budgeting models place on consumer preferences, and how these restrictions affect compensated and uncompensated price effects. See for example Gorman (1970), Blackorby, Primont and Russell (1978) and Hausman, Leonard and Zona (1994). Nevo (2000) evaluates empirically the flexibility of a multi-stage model.

[10] An early example is Baker and Bresnahan (1988). They propose a "residual" demand approach which forsakes identification of the original structural parameters in favor of amalgams of structural parameters.

form of demand reflects the way consumers evaluate product attributes. [Davis (2000) discusses these and other tradeoffs.]

Besides having to grapple with how best to restrict parameters, each of the above approaches also has to address the joint determination of prices and quantities. As in homogeneous product models, the presence of right-hand side endogenous variables raises delicate identification and estimation issues. Applied researchers can most easily address identification and estimation issues in demand and mark-up systems that are linear in the parameters. In nonlinear systems, identification and estimation questions become much more complicated. For example, the implicit "reduced form" for the nonlinear (65) and (66) system:

$$\begin{aligned} Q_1 &= k_1(Z, W, \beta; \theta, \nu, \eta), \\ &\vdots \\ Q_J &= k_J(Z, W, \beta; \theta, \nu, \eta), \\ P_1 &= l_1(Z, W, \beta; \theta, \nu, \eta), \\ &\vdots \\ P_J &= l_J(Z, W, \beta; \theta, \nu, \eta) \end{aligned} \tag{68}$$

may not be available in closed form. As argued earlier, these equations also need not have a solution or a unique solution for all values of the right-hand side variables and errors.

The value of the reduced forms in (68) is that they suggest that there are many potential instruments for prices and quantities. For example, they suggest that one may be able to use other products' attributes and cost variables as instruments. Unfortunately, most IO data sets do not have product-specific or firm-specific cost information. Even when researchers do have cost information, this information is likely to be extremely highly correlated across products. The lack of good cost covariates has forced researchers to use the attributes of other products as instruments. These studies have used both the prices of other products as instruments and the nonprice attributes of other products as instruments.

Hausman (1997) is a good example of a study that uses the prices of other products as instruments. Hausman develops and estimates a multi-stage budgeting model for varieties of breakfast cereals. Because he does not have cost data for the different cereal products, and he lacks time-varying attribute data, he resorts to using cereal prices in other markets as instruments. He justifies using these prices as follows. He first supposes that the price for brand j in market m and time period t has the form

$$\ln p_{jmt} = \delta_j \ln c_{jt} + \alpha_{jm} + \nu_{jmt}, \tag{69}$$

where c_{jt} are product-specific costs that do not vary across geographic areas, the α_{jm} are time-invariant, product-city (m) specific markups, and ν_{jmt} are idiosyncratic unobserved markups. He also assumes that the ν_{jmt} are independent across markets. This

latter assumption allows him to assert that prices in other cities are correlated with a specific market's prices and uncorrelated with the unobservable markup or cost component in prices.

From the perspective of our framework, the essential questions are: What economic assumptions motivate the pricing equation (69)? Following our homogeneous-product discussion, the pricing equation (69) could represent either a markup relation obtained from a first-order profit-maximization condition or a reduced form equation arising from the solution of a model along the lines of (68). To see the problem with the former interpretation, imagine that each manufacturer j maximizes profits of one product in each market. Suppose the firm also has constant marginal costs. If it maximizes profits by choosing quantity, then the markup equations will have the additive form

$$P_j = c_j + \tau(Q_1, \ldots, Q_J),$$

where as in (48) the $\tau(\cdot)$ function contains an own-demand derivative. We can re-express this equation as (69)

$$\ln P_j = \ln c_j - \ln\bigl(1 - \tau(Q_1, \ldots, Q_J)/P_j\bigr).$$

The question then is whether the multi-level demand function specification Hausman uses leads to the second term above having the form $\alpha_j + \nu_{jt}$, where the ν's are independent across markets. In general, his flexible demand system would not appear to lead to such a specification.

One could imagine alternatively that (69) is the reduced form obtained by simultaneously solving the first-order conditions. The problem with this view is that Hausman's flexible demand specification implies the costs of all other products should enter the reduced form. This would mean that either α_{jm} would have to be time-varying to account for the time variation in other product's costs or that c_{jt} would have to be market varying.

In principle, one might imagine adjusting the multi-level demand system or the pricing equation (69) to justify using variables from other markets as instruments. Such an exercise will require additional economic and statistical assumptions. Consider, for example, the α_{jm}'s in Equation (69). These terms appear to represent unobserved product and market-specific factors that affect markups. They might, for example, capture San Franciscans' unobserved health-conscious attitudes. These attitudes might lead San Franciscans' to have a higher demand and greater willingness to pay for organic cereals. If natural cereal makers are aware of this propensity, they might advertise more in the San Francisco market. If this advertising is not captured in the demand specification, then the demand error will be correlated with the α. Hausman recognizes this possibility by removing the brand-market α's using product-market fixed effects. Letting $\widetilde{\ }$ denote the residual prices from these regressions, his results rely on the adjusted prices:

$$\widetilde{\ln p_{jnt}} = \delta_j \widetilde{\ln c_{jt}} + \widetilde{\nu_{jnt}} \tag{70}$$

as instruments. According to Equation (69), these adjusted prices would only contain adjusted national marginal costs, and residual cost and demand factors affecting

markups. At this point, Hausman still must assume that: (1) the adjusted time-varying national marginal costs $\ln c_{jt}$ are uncorrelated with the demand and cost errors in other cities; and (2) the residual demand and cost factors affecting markups are independent of the errors in other cities.

These two assumptions have been vigorously debated by Hausman (1997) and Bresnahan (1997). Bresnahan (1997) argued that there might be common unobserved seasonal factors that affect both demand and marginal costs. To illustrate this point, Bresnahan provides an example in which a periodic national advertising campaigns translate into increased demands and markups in all markets. This results in correlation between the idiosyncratic markup terms in other markets and demand errors.[11] Whether these advertising campaigns are of great consequence for demand and price–cost estimates in a particular application is not something that can be decided in the abstract. Rather it will depend on the marketing setting and the economic behavior of the firms under study.

Our discussion so far has emphasized the strong assumptions needed to use prices in other markets as instruments. Do the same arguments apply to nonprice attributes? At first, it might seem that they might not. Similar concerns, however, can be raised about nonprice instruments. Consider, for example, the problem of trying to model airline travel demand along specific city-pairs. In such a model, the researcher might use a flight's departure time as a nonprice attribute that explains demand. The reduced form expressions in (68) suggest that besides the carrier's own departure time, measures of competing carriers' departure times could serve as instruments. But what makes' the characteristics of carriers' schedules' valid instruments? They may well not be if the carriers strategically choose departure times. For example, carriers may space their departure times to soften competition and raise fares.

If firms set nonprice attributes using information unavailable to the researcher, then we can no longer be certain that product attributes are valid instruments. In some applications, researchers have defended the use of nonprice attributes with the argument that they are "predetermined". Implicit in this defense is the claim that firms find it prohibitively expensive to change nonprice attributes in the short run during which prices are set. As a result, nonprice product characteristics can reasonably be thought of as uncorrelated with short-run unobserved demand variables. For example, a researcher modeling the annual demand for new cars might argue that the size of a car is unlikely correlated with short-run changes in demand that would affect new car prices. While this logic has some appeal, it relies on the assumption that unobserved factors affecting manufacturers' initial choices of characteristics do not persist through time. This is a question that is not easily resolved in the abstract. For this reason, models that endogenize both price and nonprice attributes continue to be an active area of research in IO.

[11] The criticism that advertising influences demand amounts to an attack on demand specifications that ignore advertising. As Hausman's empirical model does include a variable measuring whether the product is on display, the question then becomes whether the display variable captures all common promotional activity.

7.2. Micro-data models

Our discussion of the product-level demand specifications in (65) has said little about what it is that leads firms to differentiate products. One ready explanation is that firms differentiate their products in order to take advantage of heterogeneities in consumer tastes. For example, car makers regularly alter a car's styling, size, drive trains and standard features to attract particular groups of buyers. If IO economists are to understand how these models are priced and compete, it seems imperative that their demand systems explicitly recognize how consumer tastes for product attributes will affect demand at the firm level. In the language of Section 4, it seems critical that *firm-level demand* models recognize both observable and unobservable heterogeneities in *individual-level tastes*. Most neoclassical demand models, however, are ill-suited to modeling consumer heterogeneities. This is because it is unwieldy to aggregate most individual-level neoclassical demand models across consumers to obtain market or firm-level demands.

In the differentiated product literature, researchers have adopted two approaches to demand aggregation and estimation. One is to estimate individual-level demand functions for a representative sample of consumers. These demand functions are then explicitly aggregated across the representative sample to obtain market or firm demand. The second is to instead assume that consumer tastes have a particular distribution in the population. This distribution, along with individual demands, are estimated together to obtain estimates of market and firm demand.

In what follows, we explore some of the advantages and disadvantages of these two approaches. To focus our discussion, we follow recent work in IO that relies on discrete choice demand specifications. These models presume that consumers buy at most one unit of one product from among J products offered.[12] While these unitary demand models are literally applicable to only a few products, such as new car purchases, they have been used by IO economists to estimate consumer demands for a range of products.

A key distinguishing feature of these discrete choice demand models is that firms are uncertain about consumers' preferences. Firms therefore set prices on the basis of expected demand. So far, firm expectations have not figured prominently in our discussion of oligopoly models. Thus, we shall begin by showing how this type of uncertainty enters a structural oligopoly model.

Imagine there are a maximum of M_t customers at time t who might buy a car. Suppose customer i has the conditional indirect utility function for car model j of

$$U_{ijt} = U(x_{jt}, p_{jt}, \omega_{ijt}),$$

where x_{jt} is a $K \times 1$ vector of nonprice attributes of car j (such as size and horsepower), p_{jt} is the car's price, and ω_{ijt} represents consumer-level variables. Consumer i will

[12] There are continuous choice multi-product demand models. These models are better termed mixed discrete continuous models because they have to recognize that consumers rarely purchase more than a few of the many products offered. See, for example, Hanemann (1984).

buy new car j provided $U(x_{jt}, p_{jt}, \omega_{ijt}) \geqslant \max_{k \neq j} U(x_{kt}, p_{kt}, \omega_{ijt}; \theta)$. If firms knew everything about consumers' tastes, they would calculate product demand as

$$\text{Demand for product } j = \sum_{i=1}^{M_t} I(i \text{ buys new car } j), \tag{71}$$

where M_t is the number of potential new car buyers at time t and $I(Arg)$ is a zero-one indicator function that is one when Arg is true. Firms would use this demand function when it came time to set prices, and the IO researcher therefore would have to do their best at approximating this sum given the information the researcher has about the M_t consumers.

Now consider what happens when the car manufacturers do not observe some portion of ω_{ijt}. In this case, if there are no other uncertainties, the researcher would model a firm's pricing decision as based on what the firm expects demand to be:

$$\text{Expected demand} = q_{jt}^e = \sum_{i=1}^{M_t} E\Big(U(x_{jt}, p_{jt}, \omega_{ijt}) \geqslant \max_{k \neq j} U(x_{kt}, p_{kt}, \omega_{ikt}; \theta)\Big). \tag{72}$$

In this expression, E is the firm's expectation over the unobservables in ω_{ijt}. (Here, the firm is assumed to know the size of the market M_t.) The firm's expected aggregate demand for model j can equivalently be expressed as the sum of firms' probability assessments that consumers will buy model j:

$$q_{jt}^e = \sum_{i=1}^{M_t} \Pr(i \text{ buys new car } j). \tag{73}$$

This expression shows us how firms' uncertainties about their environment (i.e., their uncertainties about consumers tastes) will enter a structural model of competition. In essence, the IO researcher must now take a stand on firms' beliefs about consumers – what they know and do not know – and how this information enters consumers' tastes.

Once the researcher adopts a specific probability model for consumers' product choices, product-level demands simply are sums of consumers' purchase probabilities. These purchase probabilities and sums have the potential drawback that they may be nonlinear functions of consumer taste parameters. Despite this complication, the above formulation has one important advantage. A discrete choice model allows the researcher to model consumers' preferences over a large number of products as a function of a short list of product attributes (the x_{jt}). Thus, in contrast to the high-dimensionality of the neoclassical model, here a researcher may be able to reduce the consumers' choice problem to a choice over a few attributes.

Two crucial questions that must be answered when developing a discrete choice model are: What is it about consumer tastes that firms do not observe? And: What is a sensible model of firms' expectations? These are important questions because a researcher's inferences about price–cost margins may well be sensitive to the specification

of firms' uncertainties. In what follows, we use our framework for building structural models to evaluate two early differentiated product models. Both models estimate price–cost margins for new cars sold in the United States. The first, by Goldberg (1995), uses household-level new-car purchase data to estimate household-level purchase probabilities for different new car models. She assumes that these household-level probabilities are what firms use to determine aggregate demand. The second approach we consider is by Berry, Levinsohn and Pakes (1995). They do not have household-level data. Instead, they construct their demand system from product-level price and quantity data. Like Goldberg, they too base their demand estimates on sums of individual purchase probabilities. Unlike Goldberg, they match the parameters of this sum to realized new car market shares.

7.2.1. A household-level demand model

Goldberg's model of prices and quantities in the US new car market follows the logic of a homogeneous product competition model. Her estimation strategy is divided into three steps. In the first step, Goldberg estimates household-level demand functions. In the second, the household-level demand functions are aggregated to form estimates of firms' expected demand curves. In the third step, Goldberg uses the estimated expected demand curves to calculate firms' first-order conditions under the assumption that new car manufacturers are Bertrand–Nash competitors. From these first-order conditions, she can then estimate product-level marginal cost functions and price–cost markups for each new car model. The main novelty of Goldberg's paper is that she uses consumer-level data to estimate firms' expected new car demands. The supply side of her model, which develops price–cost markup equations, follows conventional oligopoly models, but it is computationally more difficult because the demands and derivatives for all the cars sold by a manufacturer enter the price–cost margin equation for any one new car it sells.

7.2.2. Goldberg's economic model

Goldberg's economic model treats consumers as static utility maximizers. She computes firms' expected demand as above:

$$q_{jt}^{e} = \sum_{i=1}^{M_t} \Pr(i \text{ buys new car } j). \tag{74}$$

Goldberg of course does not observe firms' expectations. The initial step of her estimation procedure therefore seeks to replace $\Pr(\cdot)$ with probability estimates from a discrete choice model. The validity of this approach hinges both on how close her discrete choice probability model is to firms' assessments and how accurately she is able to approximate the sum of probability estimates.

To estimate household probabilities, Goldberg uses data from the US Bureau of Labor Statistics Consumer Expenditure Survey (CES). This survey is a stratified random

sample of approximately 4500 to 5000 US households per quarter. By pooling data for 1983 to 1987 Goldberg is able to assemble data on roughly 32,000 households purchase decisions. In her data she observes the vehicles a household purchases and the transaction price. She augments this consumer-level data with trade information about new and used car attributes.

A critical component of her expected demand model is the list of attributes that enter consumers' utility functions. While the transactions price is clearly a relevant attribute, economics provides little guidance about what other attributes might enter consumers' utilities. Goldberg's approach is to rely on numerical measures found in car buyer guides. These measures include: horsepower, fuel economy, size, and dummy variables describing options.

In estimating the expected demands faced by new car manufacturers, Goldberg relies on the representativeness and accuracy of the Consumer Expenditure Survey. The assumption that her probability model replicates the firms' assessments of consumer behavior allows her to replace $\Pr(k \text{ buys new car } j)$ in (74) with an econometric estimate, $\hat{\Pr}(k \text{ buys new car } j)$, which is sample household k's purchase probability. The assumption that the CES sample is representative of the M_t consumers allows her to replace the sum over consumers in (75) with a weighted sum of the estimated household purchase probabilities:

$$\text{Estimated expected demand for product } j = \sum_{k=1}^{S_t} w_{kt} \hat{\Pr}(k \text{ buys new car } j), \quad (75)$$

where the w_{kt} are CES sampling weights for sample household k and S_t is the number of sample households in year t.

On the production side, Goldberg assumes that new car manufacturers maximize static expected profits by choosing a wholesale price. Unfortunately Goldberg does not observe manufacturers' wholesale prices. Instead, she observes the transactions prices consumers paid dealers. In the US, new car dealers are independent of the manufacturer. The difference between the retail transaction price and the wholesale price thus reflects the independent dealer's markup on the car. The dealer's incentives are not modeled in the paper for lack of data. Because Goldberg is modeling manufacturer's pricing decisions (and not transactions prices), Goldberg assumes that there is an exact relation between the unobserved wholesale prices and average transactions prices she computes from the CES. For example, she assumes that the wholesale price of an intermediate-size car is 75% of an average transaction price she can compute from the CES. While this assumption facilitates estimation, it is unclear exactly why it is profit-maximizing for dealers and manufacturers to behave in this way.[13]

Goldberg models manufacturers' decisions about wholesale prices as outcomes of a static Bertrand–Nash pricing game in which manufacturers maximize expected US

[13] For more discussion of automobile dealer behavior see Bresnahan and Reiss (1985).

profits. The expectation in profits is taken over the demand uncertainty in each ω_{ijt}.[14] Thus, firm f maximizes

$$\max_{p_t^{wf}} \sum_{j=1}^{n_f} \left(p_{jt}^{w} - c_{jt}\right) E\left[q_{jt}\left(p^{w}\right)\right], \tag{76}$$

where $p_t^{wf} = (p_{1t}^{wf}, \ldots, p_{n_f,t}^{wf})$ is a vector of wholesale prices, n_f is the number of new car models offered by firm f and c_{jt} is the constant marginal production cost for new car model j. The first-order conditions that characterize manufacturers' wholesale pricing decisions have the form

$$p_{jt}^{wf} q_{jt}^{e} + \sum_{k=1}^{n_f} \frac{p_{kt}^{wf} - c_{kt}}{p_{kt}^{wf}} q_{kt}^{e} \epsilon_{kjt} = 0, \tag{77}$$

where $q_{kt}^{e} = E(q_{kt})$, and $\epsilon_{kjt} = \frac{p_{jt}^{wf}}{q_{kt}^{e}} \frac{\partial q_{kt}^{e}}{\partial p_{jt}^{wf}}$ is the cross-price elasticity of expected demand. This equation shows that in order to obtain accurate estimates of the firm's price–cost margins, we need to have accurate estimates of the firms' perceived cross-price elasticities. Changes in the demand model, say by changing the model of firm uncertainty about consumer tastes, will likely change the estimated cross-price elasticities, and thus in turn estimates of price–cost markups.

Once Goldberg has estimated her demand model and obtained expressions for the cross-price elasticities, the only remaining unknowns in the firms' first-order conditions are their marginal costs, the c_{jt}. Because Goldberg has one first-order condition for each product, she can in principle solve the system of equations exactly to obtain estimates of the c_{jt} and price–cost margins.

7.2.3. The stochastic model

To estimate household purchase probabilities, Goldberg employs a nested logit discrete choice model. She assumes consumers' conditional indirect utilities have the additive form

$$U_{ijt} = U(x_{jt}, p_{jt}, \bar{\omega}_{ijt}) + v_{ijt},$$

where $\bar{\omega}_{ijt}$ are observable household and product characteristics and v_{ijt} is a generalized extreme value error. Goldberg goes on to assume that the indirect utility function is linear in unknown taste parameters, and that these taste parameters weight household characteristics, vehicle attributes and interactions of the two. The generalized extreme

[14] In principle, the firm also might be uncertain about its marginal cost of production. Goldberg can allow for this possibility only if the cost uncertainty is independent of the demand uncertainty. Otherwise, Goldberg would have to account for the covariance of demand and costs in (76).

value error assumption appears to be made because it results in simple expressions for the firms' expectations about consumer purchase behavior found in Equation (74).

The generalized extreme value error results in a nested logit model. Goldberg's choice of logit nests is consistent with but does not imply a particular sequential model of household decision making. Specifically, she expresses the probability that household k buys model j as a product of conditional logit probabilities:

$$\begin{aligned}
&\Pr(k \text{ buys new car } j)\\
&\quad = \Pr(k \text{ buys a car}) \times \Pr(k \text{ buys a new car} \mid k \text{ buys a car})\\
&\qquad \times \Pr(k \text{ buys new in segment containing } j \mid k \text{ buys a new car})\\
&\qquad \times \Pr(k \text{ buys new from } j\text{'s origin and segment} \mid k \text{ buys new}\\
&\qquad\qquad \text{in segment containing } j)\\
&\qquad \times \Pr(k \text{ buys } j \mid k \text{ buys new from } j\text{'s origin and segment}). \qquad (78)
\end{aligned}$$

This particular structure parallels a decision tree in which household k first decides whether to buy a car, then to buy new versus used, then to buy a car in j's segment (e.g., compact versus intermediate size), then whether to buy from j's manufacturer – foreign or domestic, and then to buy model j.

Goldberg appears to favor the nested logit model because she is uncomfortable with the logit model's independence of irrelevant alternatives (IIA) property. The IIA property of the conventional logit model implies that if she added a car to a consumer's choice set, it would not impact the relative odds of them buying any two cars already in the choice set. Thus, the odds of a household buying a Honda Civic relative to a Toyota Tercel are unaffected by the presence or absence of the Honda Accord. The nested logit corrects this problem by limiting the IIA property to products within a nest.

In principle, Goldberg could have chosen a different stochastic distribution for consumers' unobserved tastes, such as the multivariate normal. Goldberg makes it clear that she prefers generalized extreme value errors because they allow her to use maximum likelihood methods that directly deliver purchase probability estimates. Specifically, the nested logit model permits her to compute the right-hand side probabilities in (78) sequentially using conventional multinomial logit software. Her choice of nesting structure is important here because the IIA property holds at the household level for each new car within a nest. Changes in the nests could affect her estimates of cross-price elasticities. Unfortunately, economic theory cannot guide Goldberg's nesting structure. This ambiguity motivates Goldberg to explore at length whether her results are sensitive to alternative nesting structures.

While the independence of irrelevant alternatives applies to some household choices, it does not apply at the market demand level. This is because Goldberg interacts income and price with household characteristics. By using interactions and aggregating using household sampling weights, Goldberg insures that her product-level demand functions do not have the economically unattractive IIA structure.[15]

[15] This can be seen by examining the population odds of buying two different vehicles.

Goldberg makes two other key stochastic assumptions when she estimates her nested logit model. The first is that new car prices and nonprice attributes are independent of consumers' unobserved tastes, the ν_{ijt}. This is a critical modeling assumption, as it is possible to imagine cases where it would not hold. Suppose, for instance, that the ν_{ijt} includes consumer perceptions about a car's quality, and that firms know consumers perceptions. In this case, firms' pricing decisions will depend on the car's quality. Because Goldberg does not observe quality, her econometric specification will attribute the effects of quality to price and nonprice attributes. This results in the same endogeneity problem found in neoclassical demand models. To see the parallel, imagine that ν_{ijt} consists of a product-time fixed effect ("quality") and noise. That is, $\nu_{ijt} = \xi_{jt} + \eta_{ijt}$. Because ξ_{jt} is common to all households and known to the firm, it will appear in the aggregate demand curve

$$q^e_{jt}(\xi_{jt}) = \sum_{i=1}^{M_t} \Pr(i \text{ buys new car } j \mid \xi_{jt})$$

that the manufacturer uses when choosing wholesale prices to maximize profits. Thus, wholesale prices will depend on unobserved quality. Because Goldberg does not observe product quality, she needs to devise a strategy for removing any potential correlation between price and consumers' unobserved tastes.

The best way to account for this unobserved heterogeneity within a nested logit model would be to add behavioral equations to the model that would explain how manufacturers jointly choose price and quality. Such a formulation unfortunately complicates estimation considerably. As an alternative, Goldberg could simply assume a distribution for quality and then integrate quality out of aggregate demand using this assumed distribution. This strategy is economically unattractive, however, because one would have to recognize the unknown correlation of prices and qualities when specifying the joint distribution. What Goldberg does instead is assume that unobserved quality is perfectly explained by a short list of time-invariant product characteristics, such as the manufacturer's identity (e.g., Toyota), the country of origin (e.g., Japan) and the car's segment (e.g., compact). The assumption of time invariance allows her to use fixed effects to capture these components. The ultimate question with this strategy that cannot be easily answered is: Do these fixed effects capture all the product-specific unobservables that might introduce correlation between prices and consumers' unobserved preferences? Goldberg provides arguments to suggests that these fixed effects are adequate. In principle, if she had a dataset that contained many purchases of each model, she could include a complete set of model-specific dummy variables, and thereby control for all unobserved (to the researcher) quality differences across models.

A final stochastic component of the model pertains to manufacturers' marginal costs. The system of first-order conditions (77) exactly identifies each product's marginal costs. Following Hausman, Leonard and Zona (1994), Goldberg uses these marginal cost estimates to calculate product price–cost markups, which she finds to be somewhat on the high end of those reported in other studies.

Goldberg also is interested in assessing how marginal costs are related to vehicle characteristics and market conditions. To do this, she assumes that the implied marginal costs which she recovers depend on observable product characteristics and an unobservable according to

$$c_{jt} = c_0 + Z_{jt}\alpha + u_{jt},$$

where the Z_{jt} are observable product characteristics and u_{jt} are unobservable factors affecting costs. The error in this relation accounts for the fact that the estimated marginal costs are not perfectly explained by observables.

7.2.4. Results

If we compare Goldberg's model to homogeneous product competition and neoclassical differentiated product models, we see that Goldberg's competition model is considerably richer. Her demand system (75) admits complicated substitution patterns among products. These substitution patterns depend on the proximity of products' attributes. There are two main costs to this richness. First, she must introduce many functional form and stochastic assumptions to limit the scale and computational complexity of the model. As we argued earlier, structural modelers often must introduce assumptions to obtain results. Without these assumptions and restrictions, it would be impossible for Goldberg to estimate demand and costs, or evaluate the impact of the voluntary export restraints. She also might not be able to argue convincingly that her estimates make sense (e.g., that they imply a pure-strategy equilibrium exists or is unique).

A second cost of the richness of her model is that it becomes difficult for her to summarize exactly how each economic and stochastic assumption impacts her conclusions. For example, at the household level she maintains IIA within nests. Her utility specifications and method of aggregation, however, imply that IIA will not hold at the aggregate level. But just how much flexibility is there to the aggregate demand system and the cross-price elasticities? Questions about the role of structural assumptions such as this are very difficult to answer in complex models such as this. For this reason, Goldberg and other structural modelers must rely on sensitivity analyses to understand how their conclusions depend on their assumptions. For instance, Goldberg spends considerable time exploring whether her parameter estimates and implied markups agree with other industry sources and whether the estimates are sensitive to alternative plausible structural assumptions.

While structural researchers can in many cases evaluate the sensitivity of their estimates to specific modeling assumptions, some aspects of structure are not so easily evaluated. For example, Goldberg's model relies on the maintained assumption that the weighted sum of estimated CES sample purchase probabilities accurately measures firms' expectations about product demand. If there is something systematic about firms' expectations that her household model does not capture, then this will mean she is not solving the same first-order profit maximization problems that the firms were when

they set prices. Her reliance on this assumption is nothing new. The correct specification of demand is implicit in other papers in this area (e.g., Porter and Hausman). As we argued earlier in laying out our framework, all structural models base their inferences on functional form and stochastic assumptions that are in principle untestable. In this case, Goldberg's problem is that she does not observe firms' expectations. Consequently, when she finds that her model under-predicts total new car sales, she cannot know for sure whether this is because firms underpredicted demand or there is a problem with her econometric specification or data.[16]

7.3. A product-level demand model

Bresnahan (1981, 1987) was the first IO economist to use the discrete-choice demand model in an oligopoly equilibrium model to estimate the extent of firm market power. Bresnahan's preferences assume that consumers trade off the price of the product against a single unobserved quality index. Berry (1994) and Berry, Levinsohn and Pakes (1995) (BLP) extended Bresnahan's single-index model to allow products to be horizontally differentiated. In what follows, we describe BLP's (1995) original model and compare it to Goldberg's model and the neoclassical demand systems discussed earlier. Unlike Goldberg, Bresnahan and BLP only have access to product-level data. Specifically, they know a new car model's: unit sales, list price, and attributes. BLP, for example, have twenty years of data covering 2217 new car models. Their definition of a new car model (e.g., Ford Taurus) is rich enough to describe important dimensions along which new cars differ. Their data, however, do not capture all dimensions, such as difference in some two-door versus four-door models, and standard versus "loaded" models.

BLP use these product-level price and quantity data to draw inferences about consumer behavior and automobile manufacturers' margins. Like Goldberg, they base their demand system on a discrete choice model of consumer choices. At first this may seem odd – how can they estimate a consumer choice model with aggregate data? The answer lies in the structural assumptions that permit them to relate household decisions to product-level price and quantity data.

We can informally contrast Goldberg and BLP's approaches by comparing how they model the product demands on which firms base their pricing decisions. Recall Goldberg computes firms' expected product demands as follows:

$$q_{jt}^{e} = \sum_{i=1}^{M_t} \Pr(i \text{ buys new car } j) = \sum_{i=1}^{M_t} \Psi(p_{0t}, \ldots, p_{Jt}, x_{0t}, \ldots, x_{Jt}, \bar{\omega}_{ijt}; \theta), \quad (79)$$

where the $\Psi(P, x, \bar{\omega}_{ij}; \theta)$ are the nested logit purchase probabilities that depend on the price, p, and nonprice attributes, x, of all models. Because Goldberg only uses

[16] Goldberg's chief hypothesis is that the household CES data under-represent total sales because they do not include government, business or other institutional sales.

household-level data, there is no guarantee that when she aggregates her probability estimates to form q^e_{jt} that they will match actual aggregate US sales figures, q_{jt}.

BLP (1995) on the other hand do not have the household-level data required to estimate how household choice probabilities vary with $\bar{\omega}_{ijt}$. Instead, they treat actual sales, q_{jt}, as though it is a realization from the demand curve that the firm uses to set price. In essence, they assume $q_{jt} = q^e_{jt}$. BLP then replace the household-specific probabilities $\Pr(P, x, \bar{\omega}_{ij}; \theta)$ on the right-hand side with unconditional purchase probabilities $\mathcal{S}_j(P, x, \theta)$. They do this by assuming a distribution, $P(\bar{\omega}_{ijt}, \gamma)$, for the household variables that they do not observe. Here γ denotes a set of parameters that indexes the density. Formally, they compute the unconditional demand functions

$$q^e_{jt} = \sum_{i=1}^{M_t} \int_{\omega} \Phi(p_t, x_t, \omega; \theta)\, \mathrm{d}P(\omega; \gamma) = M_t \mathcal{S}_j(p_t, x_t; \theta, \gamma), \tag{80}$$

where $\Phi(\cdot)$ are choice probabilities. The second equality follows because by assumption the distribution of consumer variables is the same for each of the M_t households in the market for a new car. To estimate the demand parameter vector θ and distribution parameter vector γ, BLP match the model's predicted expected sales $q^e_{jt} = M_t \mathcal{S}_j$ to observed sales q_{it}. (This is the same as matching expected product shares $\mathcal{S}_j$ to realized product market shares, q_{it}/M_t.) As in Goldberg's econometric model, the economic and stochastic assumptions that go into the construction of $\Pr(\cdot)$ and $\mathcal{S}_j$ have a critical bearing on the resulting demand and markup estimates.

It is useful to reiterate the differences in data requirements and modeling assumptions between Goldberg and BLP. BLP fit their model to match aggregate market shares, where market shares are national sales divided by a hypothesized number of potential buyers at time t, M_t. Consequently, the reliability of demand estimates obtained will depend on the quality of the estimates of M_t. This in turn will impact the reliability of their cost estimates. In contrast, Goldberg fits a household-level model and does not require market-level data. But as noted earlier, this data set excludes some purchases by businesses and government agencies that are relevant to firms' pricing decisions. This could impact her cost estimates.

7.3.1. The economic model in BLP

BLP's economic model of automobile sales maintains that manufacturers sell new cars directly to consumers. Manufacturers do not price discriminate and consumers are assumed to know the prices and attributes of all new cars. There are no intertemporal considerations for either firms or consumers. In particular, there is no model of how firms choose product attributes, and consumers do not trade off prices and product attributes today with those in the future.

As before, consumer i's conditional indirect utility function for new cars has the form

$$U_{ijt} = U(x_{jt}, p_{jt}, \omega_{ijt}).$$

Consumers decide to buy at most one new car per household. There are no corporate, government or institutional sales. In contrast to Goldberg, BLP do not model the choice to buy a new versus a used car. Instead, purchases of used vehicles are grouped with the decision to purchase a hypothetical composite outside good labeled product 0. Consumers demand the outside good if they do not buy a new car. Thus, if $\sum_{j=1}^{J} q_{jt}$ is the observed number of new cars bought in year t, $q_{0t} = M_t - \sum_{j=1}^{J} q_{jt}$ is the number choosing to purchase the outside good.

The firm side of the market in BLP is similarly straightforward. Sellers know the demand functions calculated above and each others' constant marginal costs of production. Sellers maximize static profit functions by choosing the price of each model they produce. When choosing price, sellers act as Bertrand–Nash competitors, as in Goldberg.

7.3.2. The stochastic model

There are three key sets of unknowns in BLP's model: the number of consumers in each year, M_t; the distribution of consumer characteristics $\Pr(\omega; \gamma)$; and sellers' manufacturing costs. We consider each in turn.

Not knowing M_t, the overall size of the market, is a potential problem because it relates the choice probabilities described in Equation (80) to unit sales. BLP could either estimate M_t as part of their econometric model or base estimation on some observable proxy for M_t. Although the first of these approaches has reportedly been tried, few if any studies have had much success in estimating the overall size of the market. This difficulty should not be too surprising because the absence of data on the outside good means that additional assumptions will have to be introduced to identify the overall size of the market.

One way to develop intuition for the assumptions needed to estimate M_t in a general model is to consider the role M_t plays in a cross-section logit model. Specifically, suppose that utility consists of an unobserved product attribute ξ_j and an extreme value error η_{ij}:

$$U_{ij} = \xi_j + \eta_{ij}. \tag{81}$$

To obtain the unconditional purchase probabilities $\mathcal{S}_j(p, x; \theta, \delta)$ we integrate out the consumer-level unobservables

$$\mathcal{S}_j = \int_{-\infty}^{\infty} \prod_{k \neq j} F(\xi_j - \xi_k + \tau) f(\tau) \, \mathrm{d}\tau, \tag{82}$$

where $F(\xi_j - \xi_k + \tau) = \Pr(\xi_j - \xi_k + \tau > \eta_{ik})$ and $f(\cdot)$ is the density of η. The integral in (82) yields the logit probabilities

$$\mathcal{S}_j = \frac{\exp(\xi_j)}{\sum_{k=0}^{J} \exp(\xi_k)}. \tag{83}$$

The demand functions are then

$$q_j = M\mathcal{S}_j(\xi_0, \ldots, \xi_J) \tag{84}$$

or using (83)

$$\ln q_j = \ln M + \xi_j - \ln\left(\sum_{k=0}^{J} \exp(\xi_k)\right). \tag{85}$$

The demand parameters here are $\theta = (\xi_0, \xi_1, \ldots, \xi_J, M)$. As a simple counting exercise, we have J equations in J observed new vehicle quantities, and $J + 2$ unknowns, $\theta = (\xi_0, \xi_1, \ldots, \xi_J, M)$. Adding a quantity equation for the unobserved quantity of the outside good, q_0, does not change the difference between unknowns and knowns, but does allow us to collapse the log-quantity equations to

$$\ln q_j - \ln q_0 = \xi_j - \xi_0. \tag{86}$$

Since by definition $q_0 = M - \sum_{j=1}^{J} q_j$, we can rewrite the J equations as

$$\ln q_j - \ln\left(M - \sum_{j=1}^{J} q_j\right) = \xi_j - \xi_0. \tag{87}$$

In general, we require at least two restrictions on the $J+2$ unknown demand parameters $(\xi_0, \xi_1, \ldots, \xi_J, M)$ to be able to solve these J equations. Since the outside good is not observed, we can without loss of generality normalize ξ_0 to zero. This still leaves us one normalization short if M is unknown.

In their empirical work, BLP choose to fix M rather than restrict the ξ's or other parameters. Specifically, BLP assume that M_t is the total number of US households in year t. This choice has some potential shortcomings. Not all households can afford a new car. As in Goldberg, entities other than households purchase new vehicles. In principle, one could model these discrepancies by assuming that the total number of US households is a noisy measure of M_t, i.e., $\widetilde{M}_t = M_t + \Delta_t$. Substituting $\widetilde{M}_t$ into (87) with $\xi_0 = 0$ gives

$$\ln q_j - \ln\left(\widetilde{M}_t - \sum_{j=1}^{J} q_j\right) = \tilde{\xi}_j. \tag{88}$$

If we overestimate the size of the market (i.e., $\widetilde{M}_t > M_t$) then the left-hand side is smaller than it would otherwise be by the same amount for each product. This will make the average (unobserved) ξ_j seem lower, or in other words that all new cars that year are worse than average. In essence, the unobserved product qualities would act as a residual and capture both true quality differences and measurement error in the size of the market.

In actual applications, we will never know whether we have over-estimated or under-estimated M_t. This means that we will not know the direction of any bias in estimated

product qualities, the ξ_j's. While the availability of panel data might allow us to attempt a random measurement error model for M_t, in practice the nonlinearity of the demand functions in the measurement error will make it difficult to draw precise conclusions about how this measurement error impacts demand estimates. Thus, one is left with either using a proxy for M_t as though it had no error or imposing enough additional restrictions on the demand model so that M_t can be estimated.

The second set of unobservables that enter BLP's demand functions are the household variables, ω_{ijt}. Formally, BLP assume household i's indirect utility for new car j has the additive two-part structure:

$$\begin{aligned} U_{ijt} &= \overbrace{\qquad\delta_{jt}\qquad}^{} + \overbrace{\qquad\qquad\omega_{ijt}\qquad\qquad}^{} \\ &= \overbrace{x_{jt}\beta + \xi_{jt}}^{} + \overbrace{x_{jt}\nu_i + \alpha \ln(\nu_{iy} - p_{jt}) + \eta_{ijt}}^{}. \end{aligned} \tag{89}$$

The δ_{jt} includes only product attributes. For BLP it consists of a linear function of observed (x) and unobserved (ξ) product attributes. The elements of the $K \times 1$ parameter vector β are interpreted as population average marginal utilities for the observed attributes.

The ω_{ijt} contain three separate household-level terms. The familiar extreme value error term η_{ijt} allows for unobserved household-specific tastes for each model in each year. The $K \times 1$ vector of unobservables ν_i allows for the possibility that household i's marginal utilities for observed attributes differ from the population average marginal utilities (the β's). While in principle one might expect that households' marginal utilities would depend on household income and other demographic characteristics, the lack of household data forces BLP to assume that the ν_i's are independent random variables that are identically distributed in the population.[17] BLP assume these random variables are normally distributed. In addition, they assume that a household's unobserved marginal utility for attribute k is independent of their marginal utility for attribute h. The unboundedness of the support of the normal distribution implies that some households will prefer attribute k and some will have an aversion to it. Specifically, the fraction that dislike attribute k is given by $\Phi(-\beta_k/\sigma_{ik})$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function and σ_{ik} is the standard deviation of ν_{ik}.

The final stochastic component of ω is $\alpha \ln(\nu_{yi} - p_{jt})$, where α is an unknown coefficient. We use the notation ν_{yi} for income to indicate that, like the unobserved marginal utilities for the observed attributes, income also is an unobservable. The expression in the natural logarithm is the (unobserved) income the household has left if it purchases model j. BLP include $\ln(\nu_{yi} - p_{jt})$ so that they can interpret $U_{ijt}(\cdot)$ as a conditional indirect utility function. Once again they need to make some distributional assumption on the unobserved ν_{yi} in order to compute expected demand. In their empirical work BLP assume that the natural logarithm has a lognormal distribution. However, the lognormal distribution must be truncated to make the expenditure on the outside good

[17] BLP and others have explored alternatives to this structure. For example, BLP (2004) allow consumers' marginal utilities to depend on observable and unobservable household attributes.

positive. That is, they need to guarantee that $v_{yi} > p_{jt}$ for all observed and plausible counterfactual prices p_{jt}.

A final element of the preference specification is BLP's treatment of the outside good. BLP assume that the utility for the outside good has the form:

$$U_{i0t} = \alpha \ln v_{iy} + \sigma_0 \nu_{i0} + \eta_{i0t}.$$

Unobserved income enters this utility because it is the amount available to be spent on the outside good when no new car is purchased. No price enters the conditional indirect utility for the outside good because p_0 has been assumed to equal zero. The parameter σ_0 is new; it represents the standard deviation of the household's unobserved preference for the outside good, ν_{i0}. In essence, ν_{i0} increases or decreases the unobserved product qualities, the ξ_j, for household i by the same amount. By adding the same household-specific constant to the ξ's, BLP preserve households' rankings of all new cars based on their unobserved qualities, but allow households to disagree on the overall quality of new cars. To see this, suppose for simplicity that $\alpha = 0$ and $\beta = \nu_i = 0$. Utilities then are as in Equation (89) except $U_{i0} = \sigma_0 \nu_{i0} + \eta_{i0}$. The logit probabilities of purchase in (83) now have the household-specific form

$$\mathcal{S}_{ij} = \frac{\exp(\xi_j - \sigma_0 \nu_{i0})}{1 + \sum_{k=1}^{J} \exp(\xi_k - \sigma_0 \nu_{i0})}. \tag{90}$$

Thus, households with large values of ν_{0i} do not think that the quality of new cars is very high and consequently are more likely to opt for the outside good. Similarly, holding the unobserved qualities of new cars fixed (the ξ), increases in σ_0 reduce the importance of the unobserved car model qualities for purchase decisions.

7.4. More on the econometric assumptions

Now that we have provided an overview of BLP's economic and stochastic assumptions, it is useful to revisit some of them to understand further why BLP adopt these assumptions.

7.4.1. Functional form assumptions for price

A critical component of any choice model is the way in which product prices enter utility. Consider what would happen, for example, if BLP had entered (as some studies do) price as an additive function in δ_{jt} rather than in ω_{ijt}. In a standard logit choice model, with $\delta_{jt} = g(p_{jt}) + \tilde{\delta}_{jt}$, the demand equations have the form

$$\ln q_{jt} = \ln M_t + g(p_{jt}) + \tilde{\delta}_{jt} - \ln\left(1 + \sum_{k=1}^{J} \exp\bigl(g(p_{kt}) + \tilde{\delta}_{kt}\bigr)\right). \tag{91}$$

The implied own-price and cross-price elasticities for these demands are:

$$\frac{\partial \ln q_{jt}}{\partial \ln p_{kt}} = \begin{cases} \frac{\partial g(p_{jt})}{\partial p_{jt}} p_{jt}(1 - \mathcal{S}_{jt}), & k = j, \\ -\frac{\partial g(p_{jt})}{\partial p_{kt}} p_{kt}\mathcal{S}_{kt}, & k \neq j. \end{cases} \tag{92}$$

These expressions show how the extreme value error assumption and the choice of $g(\cdot)$ affect the structure of the own-price and cross-price elasticities that enter the price-markup equations. If price enters logarithmically (e.g., $g(p_{jt}) = \theta \ln p_{jt}$), then the own-price and cross-price elasticities only depend on product market shares. In this case, an increase in the price of a Jaguar would cause the demand for BMWs and Kias, which have roughly similar shares, to increase roughly the same amount, even though BMWs and Kias are hardly substitutes. To some extent, one could consider fixing this problem by changing the way price enters δ_{jt} or by interacting functions of price with other vehicle attributes. Such an approach, however, ultimately may not capture what one might expect, which is that products with similar attributes will have higher cross-price elasticities.

The use of the extreme value error can also have some other unattractive economic consequences. One consequence of the error's unbounded support is that with finite attributes, there always will be someone who will buy a product – no matter how inferior the car is to other cars. Suppose, for example, that instead of having price enter logarithmically, the function $g(p)$ is bounded above. In this case, product demands will asymptote to zero instead of intersecting the price axis. This asymptotic behavior can have an unfortunate impact on global welfare and counterfactual calculations. Petrin (2002), for example, finds that when price is entered linearly that one can obtain implausibly large estimates of the value of Minivans. Figure 1 illustrates this problem for two alternative specifications of $g(\cdot)$ using a standard logit model for shares. The demand curve labeled A assumes price enters δ as $-\lambda p$. The concave demand curve B adopts a logarithmic specification paralleling BLP, $g(p) = \lambda \ln(100 - p)$. The constant λ is selected so that each curve predicts roughly the same demand for a range of prices between 60 and 90. (One might think of this as approximating a range of data that the researcher would use to estimate λ.) Comparing the two demand curves, we can see that there would not be too much of a difference in the two models' predicted demands or local consumer surplus calculations for prices between 60 and 90. But often researchers perform calculations that rely on the shape of the demand function for all positive prices. A common example is determining the welfare gain to consumers from the introduction of a new good [e.g., Petrin (2002) and Hausman (1997)]. In this case, the properties of the demand function for the new good from the price where the demand curve intersects the vertical axis to the actual market price determine the benefits consumers derive from the existence of this good. The difference in this calculation for the two demand curves would be dramatic. For example, Demand Curve A estimates that there are many consumers with reservation prices above 100, while Demand Curve B says there are none.

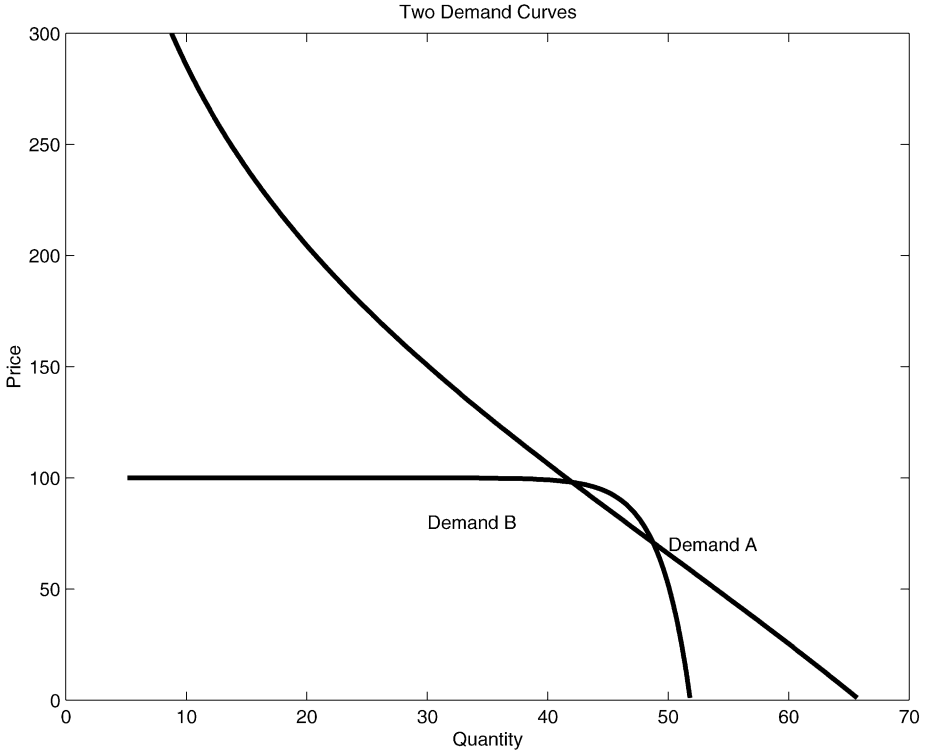

Figure 1.

7.4.2. *Distribution of consumer heterogeneity*

In their empirical work, BLP emphasize that they are uncomfortable with the IIA property of the standard logit choice model, and for this reason they add unobservable household-car attribute interactions. To gain some understanding of what these unobservables add, consider the following three good market:

- there are two types of cars available: large (LARGE $= 2$) and small (LARG$E = 1$);
- utilities for the large and small cars equal

$$U_{ij} = \beta_0 + \beta_L \text{LARGE}_j + \eta_{ij} = \delta_j + \eta_{ij};$$

  and
- the large car has 15 percent of the market, the small car 5 percent and the outside good the remaining 80 percent.

This utility specification perfectly explains the market shares. That is, we can match the observed market shares to the logit shares exactly:

$$0.15 = \exp(\beta_0 + 2\beta_L)/\big(1 + \exp(\beta_0 + \beta_L) + \exp(\beta_0 + 2\beta_L)\big),$$
$$0.05 = \exp(\beta_0 + \beta_L)/\big(1 + \exp(\beta_0 + \beta_L) + \exp(\beta_0 + 2\beta_L)\big). \tag{93}$$

A solution is: $\beta_L = 1.098$, and setting $\beta_0 = -3.871$. Although the deterministic utility specification predicts consumers prefer larger to smaller cars, the infinite support of the extreme value error ν_{ijt} results in some consumers having an idiosyncratic preference for small cars.

Now consider what happens with these data when we add heterogeneity in consumers' marginal utilities for size. In lieu of assuming a continuous distribution of marginal utilities, suppose for simplicity that there are just two types of consumers: those with a taste β_{L1} for size and those with a taste β_{L2} for size. Because we can potentially explain the three market shares with just two parameters, assume $\beta_0 = 0$. In addition, to avoid the complication of having to estimate the entire distribution of consumer preferences, suppose we know that 15 percent of consumers are of type 1 and the remaining 85 percent are type 2.

How does this two-type model explain the market share of the small car? It seems in principle that the two-type model could fit the market share data in the same way that the single type model did. Both types of consumers would have positive but different marginal utilities for vehicle size, and the unbounded support of the extreme value error would account for why some fraction of each type would buy an otherwise inferior car. To see whether this is the case, we again match the observed market shares to the logit shares:

$$0.15 = 0.15\frac{\exp(2\beta_{L1})}{1 + \exp(\beta_{L1}) + \exp(2\beta_{L1})} + 0.85\frac{\exp(2\beta_{L2})}{1 + \exp(\beta_{L2}) + \exp(2\beta_{L2})},$$
$$0.05 = 0.15\frac{\exp(\beta_{L1})}{1 + \exp(\beta_{L1}) + \exp(2\beta_{L1})} + 0.85\frac{\exp(\beta_{L2})}{1 + \exp(\beta_{L2}) + \exp(2\beta_{L2})}. \tag{94}$$

A solution is the type 1 consumers have a negative marginal utility for size ($\beta_2 = -2.829$) and the type 2 consumers have a positive marginal utility for size ($\beta_1 = 3.9836$). Thus, when consumers' marginal utilities are unconstrained, the choice model may explain the purchase of an inferior product by indicating that some consumers have negative marginal utilities for otherwise attractive attributes.

This example gets at the heart of IO economists' distinction between vertical and horizontal product differentiation models. In vertical models, consumers share similar opinions about an attribute, and thus will rank products the same. They may, however, differ in the strength of their preferences. In multi-attribute models, the relation between vertical and horizontal product differences and product rankings becomes more complex. For instance, even though consumers may all have positive marginal utilities for all attributes, they may rank products differently.

In most applications researchers will have only a few attributes that they can use to explain why consumers prefer one product over others. When there are many products

compared to attributes, a large number of products may appear "dominated" according to a pure vertical model. For example, the Volkswagen Beetle is a small car, has a small engine, slightly higher than average fuel economy, etc., and yet at times sold relatively well in the US. One way BLP's model could explain the apparent relative success of the Beetle would be to assign it a high unobserved quality, ξ. Alternatively, as we have seen above, the introduction of heterogeneous tastes can account for why consumers might prefer an otherwise "average" or "dominated" product. While the introduction of consumer heterogeneity can increase the flexibility of a discrete choice model, this increased flexibility may or may not lead to results that are economically plausible. For instance, in BLP's Table IV (p. 876), they report an estimated distribution of marginal utility for miles per dollar (MP$) across consumers that is normal with mean -0.122 and a standard deviation 1.05. This estimate implies that roughly 54 percent of consumers "dislike" fuel economy, in the sense of having a negative marginal utility of miles per dollar. For the marginal utility of horsepower divided by weight of the car (HP/Weight), the estimated distribution of marginal utility is normal with mean 2.883 and standard deviation 4.628. This implies that 27 percent of consumers dislike cars with higher values of HP/Weight. Using BLP's assumption that these marginal utilities are independent implies that 14 percent of consumers prefer cars with lower values of HP/Weight and higher values of MP$. The plausibility of these estimates of the distribution customer-level heterogeneity is an open question. However, it is important to bear in mind that the assumptions BLP make about the functional form of customer's demands, the joint distribution of customer marginal utilities and income identify the joint distribution of marginal utilities that BLP recover. In contrast, one advantage of Goldberg's approach which uses household-level data, is that the marginal utilities can be identified from the household-level purchases probabilities that she estimates.

Because inferences about consumer heterogeneity are conditional on maintained functional form assumptions, it seems imperative that some effort should go into exploring the robustness of findings to distributional assumptions. To date, there has been only a modest amount of effort along these lines [see Ackerberg and Rysman (2005), Berry (2001), Bajari and Benkard (2001, 2005) and the references therein], and much more work remains to be done. In their empirical work, BLP appear to prefer the use of normal distributions because it simplifies computations. However, their computations appear to be simplified more by their assumption that marginal utilities are independent, than their assumption of normality.

7.4.3. Unobserved "product quality"

The unobserved car attributes, the ξ_{jt}, are critical stochastic components of BLP's random utility model. Although the literature sometimes refers to the ξ_{jt} as unobserved quality, they can be any combination of product-specific unobservables that enter consumers' utility in the same way. The relevance of the ξ_{jt} is perhaps best understood by returning to the cross-section logit model where $\delta_j = \xi_j$ and $\xi_0 = 0$. In this case,

demands have the form

$$\ln q_j - \ln\left(M - \sum_{j=1}^{J} q_j\right) = \xi_j. \tag{95}$$

From this equation we see that the ξ_j act as demand "errors" that insure that the econometric choice model's predicted market shares match the observed market shares. Goldberg accounts for the presence of these ξ_j through her market segment, country-of-origin, and brand fixed effects. In BLP's model it is essential that the predicted and observed market shares match. This is because BLP's theoretical model presumes that (unconditionally) *each* consumer's decision can be represented by the same multinomial choice probabilities: $(\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_J)$. Thus, with a sample size of approximately 100 million, there should be no appreciable difference between their model's predictions and observed market shares. The only way to guarantee that there will be no difference is to have a sufficiently rich parameterization of demand. The ξ's achieve just this.

As errors, the ξ are subject to arbitrary normalizations. To understand better why normalizations are necessary, let us return to the cross section logit model. Assume that $\delta_j = x_j\beta + \xi_j$, where x_j is a $K \times 1$ vector of product attributes. Now, the J equations in (87) become

$$\ln q_j - \ln\left(M - \sum_{j=1}^{J} q_j\right) = x_j\beta + \xi_j. \tag{96}$$

Assuming M is known, we have J linear equations in $J+K$ unknowns: $(\xi_1, \ldots, \xi_J, \beta)$. We therefore require K linearly independent restrictions in order to estimate the marginal utility parameters uniquely. One choice would be to set K of the ξ's to zero. BLP instead opt to place moment restrictions on the distribution of the ξ.[18] Although they do not motivate their restrictions in any detail, the computational rationale for the restrictions is readily apparent. Specifically, BLP assume that the ξ are mean independent of the observed characteristics of new cars: $E(\xi_j \mid x_1, \ldots, x_J) = 0$. This moment condition is useful because it mimics the usual least squares moment conditions, and thus, if valid, could be used to estimate the marginal utilities (the β's) in (96). In least squares, the population moment conditions are replaced by K sample moment conditions.

While imposing the population moment condition $E(\xi_j \mid x_1, \ldots, x_J) = 0$ has a useful computational rationale, it also has nontrivial economic implications. In particular, if we view ξ as an unobserved product attribute such as product quality, then we have to wonder why it would not be correlated with observable attributes. While we can think of some attributes that might be uncorrelated, such as the number of doors on a car,

[18] In principle, BLP also could have considered other restrictions on the distribution of the ξ. For example, BLP could integrate out the population market share conditions over a distribution for the ξ_j. Such an approach is problematic when the ξ_j are correlated with observables such as price because the supply side of their model suggests a complex equilibrium relationship between price and the ξ_j.

if x_j were to include the new car's price, then there would be a clear cause for concern. The concern is one of unobserved heterogeneity – the firms observe the quality that consumers assign to cars and use this information to set price. (Intuitively, firms will set higher prices for cars with higher quality.)

BLP explicitly recognize this problem and do not include price in the list of conditioning variables $x_1, \ldots, x_J$. This means that they must introduce at least one other moment condition to estimate the price coefficient. As in the Marshallian demand case, BLP in principle have many candidate variables they can use to form moment conditions, including the attributes of other vehicles. These other attributes effectively act as "instruments" for price and any other endogenous attributes.[19]

Another question that arises in this type of study is: What guarantees that nonprice attributes are valid as instruments? This is the same issue that arose in our discussion of neoclassical demand systems. One might well imagine that car manufacturers choose attributes, such as air conditioning and size, in concert with a new car's quality (or other unobservable characteristics). If this is the case, then these attributes are no longer valid instruments. To obtain valid instruments, we would presumably need to model the determinants of product attributes.

In their empirical work, BLP base estimation on sample moment conditions involving the demand and marginal cost errors (discussed below). As can be seen from the market share expressions in Equation (80), in general it is not possible to compute closed form expressions for the δ_{jt} and ξ_{jt} that enter the population moment conditions. This means in practice that the researcher must numerically invert Equation (80) or use a fixed point algorithm to solve for the ξ_{jt}. While the integral in (80) is straightforward conceptually, it is difficult to compute in practice. As an alternative, BLP use Monte Carlo simulation methods to approximate the right-hand side integral. Specifically, they use importance sampling methods to estimate the integral in (80). They then recover the δ_{jt} using a fixed-point algorithm. From estimates of the δ_{jt}, the ξ_{jt} can be recovered from the residuals of an instrumental variable regression of δ_{jt} on product attributes.

7.4.4. Cost function specifications

To this point, we have said little about the cost side. In principle, one could estimate the demand parameters without using information from the supply side. BLP appear to add the supply side for at least two reasons. First, it contributes variables that can be used in the orthogonality conditions that identify the demand parameters. Specifically, their cost-side model contributes two additional instruments (a time trend and miles per gallon). Following the approach discussed above for constructing demand error instruments, BLP now have 21 (seven instruments times 3) sample moment conditions

[19] For example in the cross section logit model we can replace the moment condition $E(\xi_j \mid p_j) = 0$ with $E(\xi_j \mid x_{k1}) = 0$, where x_{k1} is an exogenous characteristic of car k. This again gives us K moment equations. The resulting estimator is indirect least squares, in which x_{k1} serves as an instrument for price.

for the cost-side error.[20] Second, by having a supply side model, they can study how manufacturers' marginal costs seem to vary with a model's attributes.

The stochastic specification of the cost-side is fairly straightforward. Sellers equate the marginal revenues for each model with the constant marginal costs of producing that model. The researcher estimates sellers' marginal revenues by differentiating the market share functions. As in other oligopoly models, BLP decompose product marginal cost into an observable and an unobservable component. Specifically, they assume that the natural logarithm of marginal costs depends linearly on a set of cost variables and an additive error. This error is also used to form moment conditions under the assumption that its mean does not depend on new car attributes or cost variables.

7.5. Summary

BLP report estimates for several demand models. They provide elasticity and markup estimates for different new car models. They argue that these estimates roughly accord with intuition. They also make a case for their unobserved heterogeneity specification. Because of the complexity of their model, it is harder for the authors to provide a sense for how their various maintained assumptions impact their results. For instance, the markups are predicated on the Bertrand–Nash assumption, the choice of instruments, the attribute exogeneity restrictions, the stationarity and commonality of unobserved product attributes. Subsequent work, including work by BLP individually and jointly has sought to relax some of these restrictions.[21] Ongoing work by others is exploring the consequences of other assumptions in these models, and we leave it to others to survey this work.[22]

In concluding this section on differentiated product demand estimation, we want to come back to some of the themes of our structural estimation framework. Previously we emphasized that researchers should evaluate structural models in part by how well the economic and statistical assumptions match the economic environment being studied. Differentiated product models pose an interesting challenge in this regard, both because they are difficult to formulate and because data limitations often limit the flexibility that one can allow in any particular modeling format. At present, there are few standards, other than crude sanity checks, that researchers can use to compare the wide array of assumptions and estimation procedures in use. For example, to date researchers have used both neoclassical demand and discrete choice models to estimate price elasticities and markups for ready-to-eat cereal products. Ready-to-eat cereal products would hardly seem to fit the single purchase assumption of current discrete choice models. Neoclassical models suffer from their reliance in representative-agent formulations. There also

[20] Because of near collinearity concerns, they drop two of these moment conditions in estimation. That is, they base estimation on the 5 times 3 (= 15) demand instruments plus 2 times 3 (= 6) cost instruments less two demand-side instruments.

[21] For example, Berry (2001) and Berry, Levinsohn and Pakes (2004).

[22] For example, Ackerberg and Rysman (2005), Bajari and Benkard (2001), and Bajari and Benkard (2005).

have been few attempts to date made to investigate the finite sample or asymptotic performance of different estimation procedures.[23]

8. Games with incomplete information: Auctions

Over the last thirty years, economic theorists have explored a variety of game-theoretic models in which private or asymmetric information impacts economic behavior. Examples include adverse selection, contracting and auction models. In these models, agents have private information about their "type" (e.g., productivity, health status, or valuation) and general information about the joint distribution of other agents' types. Agents may also face general uncertainty about their market environment (e.g., uncertainty over prices or aggregate productivity). Within this environment, agents use their private information strategically. The econometrician typically does not know agents' private information, market uncertainties, or the distribution of agents' private information. Thus, structural models of privately-informed decisions must take into account not only unobserved private information, but also how agents' actions are influenced by private information.

Our goal here is to illustrate how the framework in Section 4 can be used to compare different econometric models of privately informed agents. Much of this discussion focuses on auction models. Auctions have recently received enormous attention in IO, and continue to serve as a proving ground for empirical models of privately informed agents. We next discuss empirical models of regulated firms' decisions and regulator behavior. These models share similarities with auction models, but also pose special modeling issues.

8.1. Auctions overview

Our discussion of inter-firm competition models emphasized that it is economic theory that allows one to move from estimates of the conditional joint density of prices and quantities, $f(P, Q \mid X, Z)$, to statements about firms' demands, firms' costs and competition. This same principle applies to models with privately informed agents – absent economic assumptions, nothing can be said about agents' behavior or their private information.

In auction studies, economists usually know:

1. each auction's format;
2. the winning bid, and possibly all bids: $B = (b_1, b_2, \ldots, b_N)$;

[23] Indeed, with panel data on products, where new products are being introduced and old ones abandoned, it is unclear what would constitute a large sample argument for consistency or efficiency. See, however, Berry, Linton and Pakes (2004).

3. item-specific or auction-specific information X (e.g., number of potential bidders, reservation price, size, quality, date of auction); and
4. bidder-specific information, $Z = (z_1, z_2, \ldots, z_N)$ (e.g., bidders' identities and size).

In ideal applications, the economist has complete information on (B_i, X_i, Z_i) for a large number $(i = 1, \ldots, I)$ of related auctions. Thus, the number of bidders and potential bidders is known; there are no missing bids; and there are no X_i or Z_i that the bidders observe that the econometrician does not. Absent an economic model of the auction, the best an empiricist can do with these ideal data is recover a consistent estimate of $g(B_i \mid Z_i, X_i)$ – the conditional density of bids given bidder and auction characteristics.

The conditional density $g(B_i \mid Z_i, X_i)$ is a statistical object. Its dimensionality depends on the number and identity of bidders, and on the variables in X_i and Z_i. The dimension of $g(\cdot)$ is critical because in order to estimate $g(\cdot)$ nonparametrically and precisely, a researcher will need a large sample of similar auctions. The auctions must be similar in the sense that they have the same number of bidders and the same X and Z variables. If this is not the case, then the researcher must divide the sample so as to estimate a separate nonparametric function for each combination of bidders, and each set of X and Z variables. For this reason, and because auction data are rarely ideal, empiricists typically do not estimate $g(B_i \mid Z_i, X_i)$ nonparametrically. Instead, they use economic theory to place considerable structure on $g(B_i \mid Z_i, X_i)$.

In what follows we first describe how, with the aid of economic theory, one can recover economic objects from nonparametric estimates of $g(B_i \mid Z_i, X_i)$ (or objects that can be derived from the density). This discussion illustrates how differing combinations of economic and statistical assumptions can be used to identify economic constructs. Because in practice it may be difficult or impossible to estimate $g(B_i \mid Z_i, X_i)$ precisely, we next discuss why one may want to assume a parametric form for $g(B_i \mid Z_i, X_i)$.

We begin our discussion of auction models by observing that there is a strong similarity between the first-order conditions estimated in homogeneous-product oligopoly models (discussed in Sections 5 and 6) and the first-order conditions of auction models. The two types of models employ substantially different stochastic assumptions however. In structural oligopoly models, the stochastic structure typically comes from the first, second and fourth sources of error described in Section 4 – namely, researcher uncertainty about the economic environment, firm uncertainty about consumers, and measurement error. By contrast, in most auction models, the stochastic structure rests *exclusively* on the second source of model error – "agent uncertainty about the economic environment" – and that uncertainty affects strategic interactions. Understanding the ramifications of these different stochastic specifications is key to understanding how structural modelers go about recovering agents' unobserved private information from data.

In auctions, we shall see it is also important to distinguish between two types of "agent uncertainty". One is private information. In this case, bidders know something that directly affects their probability of winning. The bidders are uncertain, however, about the other bidders' private information. The other type of uncertainty is common

to all bidders. In this case, bidders do not know the "common" value of the auctioned item. (They also may have different, privately held opinions about the value of the item.)

Auction models differ not only in what agents know before they bid, but also according to what they assume about whether one bidder's information is useful to another bidder. In the simplest models, agents' private information is independently distributed and useless to other agents. In more general settings, nonnegative correlations or "affiliation" among private valuations may allow bidders to use other bidders' behavior to infer something about the unknown value of the item being auctioned. As we shall see, relationships among bidders' information can have an important bearing on what a researcher can recover from auction bids.

8.1.1. Descriptive models

IO economists have devoted substantial attention recently to analyzing auction bid data. Hendricks and Paarsch (1995), Laffont (1997), and Hendricks and Porter (in press) provide excellent introductions and surveys of empirical research on auctions. Prior to the early 1990s, empirical research on auctions largely used regressions and other statistical techniques to describe how bids, or bid summary statistics, varied with auction-specific and bidder-specific variables. Of particular interest was the effect that the number of bidders had on the level and dispersion of bids, as the number of bidders was seen to be related to the extent of competition.

The results of many of these descriptive studies were difficult to interpret. This was because it was often unclear how the data or methods in these studies disentangled differences in bids due to: observable or unobservable characteristics of bidders; strategic considerations; or simply differences in bidders' beliefs about other bidders' valuations. These problems were perhaps due to the generality in which some auction models were originally cast. Additionally, these theories were not originally developed to place testable restrictions on bid distributions. As auction theories were refined and extended in the late 1970s and 1980s, empiricists began to find the theory more useful for comparing bids from different auctions and evaluating bid summary statistics.

Hendricks and Porter (1988) provide a compelling example of how empirical researchers adapted these new auction models to data. Hendricks and Porter used bids from US government offshore oil and gas lease auctions to study the effect that the presence of more-informed bidders had on the distribution of bids. In their data, they identified more-informed bidders as those bidders who owned tracts adjacent to the auctioned tract. Their logic is that, because geologic formations with oil and gas often extend over large areas, exploration activities on adjacent tracts are likely to confer an informational advantage. To develop hypotheses, Hendricks and Porter devised a theoretical model of how less-informed bidders will behave in the presence of a single more-informed bidder. In their model, there is common uncertainty about the future value of the auctioned tract (say because the future price of oil and the amount of resources in the ground are unknown). Their theoretical model yields an equilibrium in

which the less-informed bidders use mixed strategies and the more-informed firm uses a pure strategy. From this model, they are able to derive properties of the distribution of the maximum bid by a less-informed bidder. They compare this distribution to an *ex ante* distribution of informed bids. The two differ in several ways, including the probability that there will be no bid or a bid at the reservation price. They also derive results for the probability that the more-informed bidder will win and the profits of more-informed and less-informed bidders.

In their empirical work, Hendricks and Porter account for several ways in which lease auctions differ from the auctions in their theoretical model. First, their model assumes the presence of one informed bidder, but in their data there can be multiple informed bidders. Second, their results are cast in terms of the distribution of the maximum bid by a less-informed bidder. Unfortunately, they do not know the exact structure of this distribution. These realities lead Hendricks and Porter to estimate a flexible parametric joint density of the maximum bid submitted by the more-informed (B_M) and maximum bid submitted by the less-informed (B_L) bidders. They use these estimated densities to examine certain predictions of their theoretical model. Mapping what they did to our notation, Hendricks and Porter cannot estimate and test restrictions on $g(B_i \mid Z_i, X_i)$, but they can estimate a joint density for two elements of B_i, the maximum bids of the more-informed, B_M, and less-informed, B_L, bidders.

Another practical reality is that the government sets reserve prices (or minimum bids) in these auctions. While Hendricks and Porter argue that the presence of reserve prices does not affect the equilibrium bid functions, as a practical matter Hendricks and Porter never observe more-informed and/or less-informed bids below the reserve price. That is, the reserve prices truncate the conditional density of $g(B_i \mid Z_i, X_i)$. This leads Hendricks and Porter to model the truncated distribution of maximum bids. Specifically, they assume that absent truncation, the joint distribution of B_M and B_L follows a bivariate lognormal distribution. To handle reserve prices, they work with scaled bids: $(y_{Mk}, y_{Lk})'$, where they assume

$$y_{ik} = \ln(B_{ik}/R_k) = \left(X_k' Z_{ik}'\right)\theta_i + \epsilon_{ik},$$

$i = (M, L)$, R_k is the reserve price for auction k, and $(\epsilon_{Mk}, \epsilon_{Lk})'$ are independent and identically distributed normal random errors. The variables in X_k and Z_{ik} contain tract and sometimes bidder-specific information for each auction.

The presence of the reserve price means that Hendricks and Porter only observe the y_{ik} when they are greater than or equal to zero. This type of truncation can readily be accounted for in a maximum likelihood setting using tobit-like models. In their empirical work, they develop a likelihood-based model for the scaled bids (y_{Mt}, y_{Lt}) that takes into account truncation and correlation in the bid unobservables $(\epsilon_{Mt}, \epsilon_{Lt})'$. With this amended model, they test which elements of X and Z enter the joint density of (y_{Mt}, y_{Lt}). They find a variety of results supporting their theoretical model. For instance, conditional on a set of auction-specific observables (X), the participation and bidding decisions of informed firms are more highly correlated with measures of *ex post* tract value.

8.1.2. Structural models

Hendricks and Porter's paper illustrates many of the important challenges that structural modelers face in trying to match theoretical auction models to data. Their paper also illustrates how features of the auction, such as reserve prices, may require the econometrician to make compromises. Before we can begin to evaluate different structural econometric models of auctions, we first describe the economic objects structural modelers seek to recover from auction data. After describing these economic primitives, we turn to describing various economic and statistical assumptions that have been used to recover them.

The primary goal of most structural econometric models of auctions is to recover estimates of:

1. bidders' utilities $U = (u_1, \ldots, u_N)$ (or the joint density $f_U(U)$ of these utilities); and
2. information about the uncertainties bidders face.

In single-unit auctions, bidders are modeled as receiving a nonzero utility from winning that depends on the price bidder j paid, P_j. Depending on the type of auction being modeled, bidders' utilities from winning may also depend on unobservables, such as the *ex post* value of the auctioned item. In order to proceed, the researcher thus must make some assumption about individual risk preferences. Most models assume bidders are risk neutral. In the risk neutral case, bidder j's utility can then be modeled as the difference between the *ex post* value for the object and the price the winner pays: $u_j = v_j - P_j$.

There are several critical things to note about bidders' utilities. First, it is the price paid that enters the utility function. Depending on the auction rules, there can be a difference between the amount bidder j bids, B_j, and the amount they pay, P_j. For example, in a second-price (Vickery) purchase auction, the winner pays the second-highest price, which is less than or equal to what they bid.[24] Second, as we mentioned earlier, there can be a difference in bidder j's *ex ante* and *ex post* private assessment of the value of the item. When there is no difference between bidder j's *ex ante* and *ex post* private assessment of the value of the item, we have a private values (PV) model. In this case, the v_j and their joint density, $f(v_1, \ldots, v_N) = f(V)$, are direct objects of interest. When there is a difference between bidder j's *ex ante* and *ex post* private assessment of the value of the item, this is modeled as being due to "common values", v. These common values are unobserved by the bidders *ex ante*, but known *ex post*. To account for differences in bids with common values, the bidders are assumed to possess private information or signals, s_j. These signals are assumed generated from a distribution that is conditioned on the *ex post* common values v. Thus, in a common values setting, the

[24] We use the term purchase auction to refer to auctions where higher bids increase the likelihood of winning. Conversely, procurement auctions are auctions in which lower bids increase the chances of winning.

economic objects of interest are the signals $S = (s_1, \ldots, s_N)$, their joint conditional density $f(S \mid v)$, the common values v, and the marginal density of the common values $f_v(v)$.

To summarize our discussion of auction models: there are three main dimensions along which existing auction models differ:

1. Bidders are uncertain about the *ex post* value of the item.
2. Bidders' private information signals are correlated.
3. Bidders are symmetric in their uncertainties about other bidders' private information.

In an auction where bidders are symmetric, we can summarize the major types of auction models and their primitives in a two-by-two table. Table 1 summarizes the two major differences in the way theorists and empiricists have approached modeling auction bids. The first row of each cell gives the acronym describing the auction; the second and third rows give the information and valuation objects, and the related density functions, that a structural modeler seeks to recover.

Our characterization of the affiliated values (AV) model follows Milgrom and Weber (1982) and McAfee and McMillan (1987). In an AV model, bidders receive private signals $S = (s_1, \ldots, s_N)$ about an item's value and there are also common unknown components v. *Ex ante*, each bidder j is uncertain about the value of the item. Bidders' utilities are (symmetric) functions of the common components v and all bidders' private information signals, S. That is, $v_j = V(s_j, S_j, v)$, where S_j contains all signals but bidder j's. In these models, bidders' private valuations and the common components are assumed affiliated – which loosely means that if a subset of them are large, it is likely the remainder are large.[25] Because the equilibria of affiliated values (AV) models are usually very difficult to characterize, there have been few attempts to estimate general affiliated value models.

Table 1
Private information (conditionally) independent

Uncertainty in final value	YES	NO
YES	PCV $f_v(v)$ $f_{s\mid v}(s_j \mid v)$ $s_1, \ldots, s_N, v$	AV $f(S, v)$ $s_1, \ldots, s_N, v$
NO	IPV $f_s(s_j)$ $s_1, \ldots, s_N$	APV $f_S(S)$ $s_1, \ldots, s_N$

[25] See Milgrom and Weber (1982) for a more complete discussion and references.

IO economists have instead focused on estimating the special cases of the AV model described in the remaining three cells. The bottom row of the table describes two private values (PV) models. In a PV model there is no uncertainty to the bidder's valuation because bidder j observes s_j prior to the auction and thus knows v_j. Bidder j still faces uncertainty in a PV auction, however, because other bidders' valuations are unknown. In an asymmetric independent private values (IPV) model, bidders' presume that the other bidders' values are independently drawn from the marginal densities $f_j(s_j)$. In an affiliated private values (APV) model, nonnegative correlation is allowed. When the bidders share the same beliefs about each others' private valuations, we can represent the density of valuations in a symmetric APV model by $f(s_1, \ldots, s_N)$.

Another special case of interest is a pure common values (PCV) model. In contrast to the private values model, in a PCV model, bidders do not know the value of the item before they bid. All bidders, however, will *ex post* value the item the same. Thus, it is as though there is a single common component v and $V_j(S, v) = V_k(S, v) = v$ for all signals. Such a situation might characterize a situation where bidders are purchasing an item for later resale. To calculate the expected value of winning in a PCV auction, the researcher requires assumptions about the joint distribution of the known signals and the *ex post* value. To facilitate calculations, the usual assumptions are that there is a commonly known prior distribution for v, $f_v(v)$ and that bidders' private information conditional on the signal are (symmetrically) conditionally independent – i.e., $f_{S|v}(S \mid v) = \prod_{j=1}^{N} f_{s|v}(s_j \mid v)$.

We now discuss situations where one can recover the objects listed in this table. The standard approach to developing a structural auction model is to derive equilibrium bid functions for each bidder given each bidder's utility function, the bidder's private signal, other bidders' strategies and the bidder's beliefs about the other bidders' signals. Provided these Bayesian–Nash bid functions are increasing in the unobservable private information and any common values, the empiricist can potentially recover estimates of the unobservables. That is (in a slight abuse of notation), the structural modeler hopes to relate observed data on bids in auction i, B_i, to equilibrium bid function equations: $B_1(s_{1i}), \ldots, B_N(s_{Ni})$. While our notation suggests that the equilibrium bid function for bidder j, $B_j(s_j)$ only depends on the bidder's private information s_j, the equilibrium function also depends on the distribution of private information and common values, $F(S, v)$. This dependence means that in practice we cannot determine the specific form of $B_j(s_j)$ without either (a) making a parametric assumption about $F(S, v)$, or (b) using an estimate $g(B_i \mid Z_i, X_i)$ to recover information on the form of $F(S, v)$.

In nearly all empirical auction models, the process of drawing inferences about the objects in the above table is facilitated by the stark assumption that the only source of error in auction bids are S and v. That is, most empirical models of auctions do not allow for measurement errors in bids or unobserved heterogeneity in the valuation distribution across auctions.[26]

[26] There are exceptions. Paarsch (1992) attempts to model unobserved heterogeneity in timber auctions. Krasnokutskaya (2002) models unobserved heterogeneity in highway procurement auctions.

8.1.3. Nonparametric identification and estimation

Recently, structural modelers have devoted substantial energy to the problem of flexibly estimating the joint density of private information and a single common value component – $f(S, v)$. These efforts reflect the practical reality that the researcher rarely knows *ex ante* what specific $f(S, v)$ bidders used. While this ignorance might seem to favor the researcher estimating general nonparametric density functions for affiliated random variables, such models have proven computationally impractical. This has led researchers to focus on establishing nonparametric identification and estimation results for the special cases described in the previous table.

8.1.3.1. Private values auctions Symmetric independent private values auctions present the simplest identification issues. In a symmetric IPV model, the researcher seeks to recover an estimate of the marginal density of private information $f(s_j)$ (or equivalently, $f(v_j)$) from bid data. The main result in this literature is that data on winning bids are sufficient to nonparametrically identify $f(v_j)$ and estimate v_j. To gain an intuitive understanding of what is involved in deriving nonparametric identification results for private information models, it is useful to begin by considering what happens when there is no private information. By ignoring agent uncertainty and agent beliefs, we can isolate the effect that the econometrician's uncertainty has on inferences. To do this, we compare two procurement auctions.

The following example auction draws an analogy between the problem of how to estimate firm costs in a Bertrand oligopoly setting and the problem of how to estimate the distribution of private values in an IPV setting.

EXAMPLE 10. In a symmetric IPV procurement auction, the bidders' valuations (or in this case costs) are drawn independently from the same marginal distribution $f(c_j)$. Each bidder j only gets to observe their cost c_j. Suppose that each bidder can observe all bidders' costs, $C = (c_1, \ldots, c_N)$ so that each bidder knows the identity of the lowest cost bidder. Because no bidder will find it profitable to bid less than his cost, it is easy to see that in a Nash equilibrium the lowest-cost bidder will win the auction by bidding (slightly less than) the second-lowest cost.

This equilibrium is analogous to what would happen in a homogeneous-product Bertrand oligopoly. In a homogeneous-product Bertrand market where firms have different constant marginal costs, the firm with the lowest marginal cost will end up supplying the entire market at a price equal to the marginal cost of the second-lowest cost firm.

Now consider what an economist could learn by observing equilibrium prices in a set of Bertrand markets. Because the economist knows the equilibrium price equals the marginal cost of the second-lowest cost competitor, they can use a random sample of market prices to estimate the density, $f(c_{[2:N]} \mid X, Z)$, of the second-lowest marginal cost. Moreover, as we shall see shortly, it will be possible under certain assumptions for the economist to recover the density of mar-

ginal costs, $f(c \mid X, Z)$, from $f(c_{[2:N]} \mid X, Z)$. Thus, this example suggests how an economist might recover information about bidders' valuations in an IPV auction *from only data on winning bids*. The key simplifying assumption, which we shortly relax, is that we assumed that the Bertrand competitors were not privately informed about their costs. This makes the solution of the Bertrand and IPV auction very simple in that all but the winning bidder have an incentive to bid their true costs.

Before introducing private information among bidders, it is useful to explore what would happen if the economist had more data than just equilibrium price. For example, suppose that the Bertrand market followed the format of an English button auction. In a descending-price English button auction, all bidders start the auction facing each other with their fingers depressing buttons. The seller then announces continuously lower prices starting from a very high price. Bidders drop out of the auction when they remove their fingers from the buttons. The analogy in a Bertrand market would have prices start out at very high levels with all firms being willing to supply the market. As firms continuously undercut one another, firms would drop out once the price fell to their marginal cost. This process would continue until price hit the marginal cost of the firm with the second-lowest marginal cost. At this point, the firm with the second-lowest marginal cost would drop out and the firm with the lowest cost would supply the entire market at this price. By observing the prices at which all firms dropped out, the economist could directly infer the marginal costs of all but the most efficient firm. Thus, the economist could use the drop-out prices to improve their estimates of the density of marginal costs $f(c_j)$.

This next example considers the effect that correlation among private values has on inferences made from bid data. To do this we compare an APV model to a homogeneous-product, quantity-setting oligopoly model. Again we assume that the oligopoly firms' and bidders' costs are known. Later we will draw a more formal analogy between this oligopoly example and a PV auction.

EXAMPLE 11. Consider an N-firm homogeneous product oligopoly in which firms' constant marginal costs are drawn independently from the joint density $f(c_1, c_2, \ldots, c_N) = f(C)$. Let $P(Q)$ denote the inverse market demand curve, q_i the output of firm i, and let $Q = \sum_{i=1}^{N} q_i$ denote industry output. Assume, as in the previous example, that the suppliers observe C, the vector of marginal costs, before they choose quantity to maximize profits (given the quantity choices of their competitors). The profits of each firm are: $\pi_i(Q) = (P(Q) - c_i)q_i$. The optimal Nash equilibrium quantities solve the N first-order conditions:

$$P = c_i - \frac{\partial P(Q)}{\partial Q} q_i. \tag{97}$$

As we shall see shortly, these first-order conditions closely parallel the first-order conditions that determine equilibrium bids in private value auctions.

Using the implicit function theorem, we can solve these equations for quantities as a function of all firms' costs. Similarly we can use a change-of-variables formula to derive the joint density of $(q_1, q_2, \ldots, q_N)$ from the joint density of $(c_1, c_2, \ldots, c_N)$. Both of these operations require a nonvanishing Jacobian, which amounts to an identification condition for obtaining the joint density of firms costs, $f(c_1, c_2, \ldots, c_N)$, from the joint density of $(q_1, q_2, \ldots, q_N)$. Analogously, in an affiliated private values auction model, there are a set of first-order conditions that relate the privately-observed costs and assumed joint distribution of costs to the optimal bids. [See, for example, Li, Perrigne and Vuong (2002).] By observing a large sample of independent and identical auctions, one could construct an estimate of the joint distribution of the equilibrium bids, $g(B)$. The researcher could then substitute this estimate into the first-order conditions and (provided certain technical conditions are satisfied) use it to recover estimates of the unobserved costs.

Although the above discussion is heuristic, it makes clear that identification hinges on having sufficient similarities in the sampled auctions.[27] As soon as the sampled auctions differ in observable or other unobservable ways, there may be no practical way by which the researcher can reliably recover $f(c_1, c_2, \ldots, c_N)$ from the observed bids.

This point is perhaps easier to appreciate by considering how researchers have estimated the distribution of valuations nonparametrically in an IPV setting. Guerre, Perrigne and Vuong (2000) were the first to devise techniques for recovering a consistent estimate of the underlying distribution of IPV model valuations without making specific parametric assumptions. They model the behavior of N *ex ante* identical risk neutral bidders in a first-price auction. In the case of a procurement auction, this amounts to bidder j maximizing the expected profits

$$E\big[\pi_j(b_1, b_2, \ldots, b_N)\big] = (b_j - c_j)\Pr(b_k > b_j,\ \forall k \neq j \mid c_j), \tag{98}$$

by setting

$$b_j = c_j - \Pr(b_k > b_j,\ \forall k \neq j \mid c_j)\left(\frac{\partial \Pr(b_k \geqslant b_j,\ \forall k \neq j)}{\partial b_j}\right)^{-1}. \tag{99}$$

Here $\Pr(b_k \geqslant b_j,\ \forall k \neq j)$ is the probability that supplier j wins with a low bid of b_j and c_j is bidder j's private cost.

In the symmetric IPV case, the equilibrium bid function simplifies to

$$b_j = \beta(c_j \mid F, N) = c_j + \frac{\int_{c_j}^{\infty}[1 - F(\tau)]^{N-1}\,\mathrm{d}\tau}{[1 - F(c_j)]^{N-1}}, \tag{100}$$

where here we use $\beta(\cdot)$ to denote the equilibrium bid function, $F(c_j)$ is the distribution function of private cost (value) for the item being auctioned. This expression relates the

[27] For illustrative discussions of identification issues see Laffont and Vuong (1996), Guerre, Perrigne and Vuong (2000), and Athey and Haile (2002).

equilibrium bids explicitly to the bidder j's own cost c_j and the distribution function $F(\cdot)$ of all bidders' marginal costs. Because by assumption this expression holds for each of the N bidders across each of the I auctions, the researcher could construct and compare separate estimates of $F(c_j)$ from different random collections of observed bids.

Guerre, Perrigne and Vuong (2000, p. 529) describe a related nonparametric procedure as follows. Assuming no measurement error in the bid data, a straightforward application of the change-of-variables formula yields an expression for the density of each bid b:

$$g(b) = \frac{f(\hat{c})}{|\beta'(\hat{c})|} = \frac{f(\beta^{-1}(b))}{|\beta'(\beta^{-1}(b))|} \tag{101}$$

where $\hat{c} = \beta^{-1}(b)$ is the inverse of the equilibrium bid function, $b = \beta(\hat{c})$, $\beta'(\cdot)$ is the first derivative of $b = \beta(\hat{c})$, and $f(c)$ is the density associated with $F(c)$. Thus, $\hat{c}_j = \beta^{-1}(b_j)$ is the private cost given the observed bid b_j. To apply this formula, they require that the bid function be strictly monotone.

Equation (101) relates the density of observed bids to the unknown density of private costs, apart from the derivative of the equilibrium bid function in the denominator. By differentiating (100) one can obtain an expression for this derivative. Using this expression to substitute out the integral in (100), we obtain

$$\beta(c_j \mid F, n) = c_j + \frac{\beta'(c_j)[1 - F(c_j)]}{(N-1)f(c_j)}. \tag{102}$$

Substituting (101) into this expression and making use of $G(b_j) = F[\beta^{-1}(b_j)]$ gives

$$c_j = b_j - \frac{1}{N-1}\frac{1 - G(b_j)}{g(b_j)}. \tag{103}$$

Here $G(b)$ is the distribution of bids, $g(b)$ is the density of bids, and N is equal to the number of bidders. Thus, to recover the unobserved private costs on the left-hand side, the researcher only requires estimates of the distribution function and density function of bids. Under the assumption that there are no observable or other unobservable differences across auctions, and that $G(\cdot)$ and $g(\cdot)$ are the same across auctions, the researcher can pool data on all bids to estimate $G(\cdot)$ and $g(\cdot)$ nonparametrically. From (103), the researcher can estimate c_j. Once the researcher has estimates of the c_j, nonparametric smoothing techniques can again be used to produce an estimate of the density $f(c)$ or distribution $F(c)$ of private costs.

This same strategy can be used to estimate the density of private values nonparametrically if the researcher only observes winning bids. In this case, Equation (101) must be changed to account for the fact that the winning bid in an IPV procurement auction is that of the lowest-cost bidder. Because the winning bidder has cost $c_{(1:N)}$, the density

of the winning bid b_w is

$$h(b_w) = \frac{\bar{g}(\beta^{-1}(b_w))}{\beta'(\beta^{-1}(b_w))}, \quad \text{where } \bar{g}(z) = N\big[1 - G(z)\big]^{N-1} g(z), \tag{104}$$

and $z = c_{(1:N)}$.

The strength of this nonparametric approach is that it does not require parametric assumptions about unobserved valuations. To see why this flexibility is important economically, it is useful to compare Equation (99)

$$b_j = c_j - \Pr(b_k > b_j,\ \forall k \neq j \mid c_j)\left(\frac{\partial \Pr(b_k \geqslant b_j,\ \forall k \neq j)}{\partial b_j}\right)^{-1} \tag{105}$$

to the standard oligopoly mark-up equation

$$P = c_i - \frac{\partial P(Q)}{\partial Q} q_i.$$

In both equations, the second term on the right-hand side determines the markup over marginal cost. The numerator of Equation (105) is analogous to q_i, the quantity sold. The denominator is the decrease in the probability of winning the auctioned item with an increase in the bid, which is analogous to the decrease in quantity with an increase in price. Just as it was important in oligopoly models to use a demand model that yielded flexible demand elasticities, so too it is important to have a distribution function $F(\cdot)$ that yields flexible bid mark-ups.

There are of course costs to estimating $G(\cdot)$ and $F(\cdot)$ flexibly using nonparametric methods. Chief among them is that the researcher will require data on a large number of similar auctions. In practice the researcher may not be able to reliably estimate $F(\cdot)$ when there are more than a few dimensions of observable auction (X) or bidder (Z) heterogeneities. Moreover, reserve prices introduce the similar truncation issues to those in Hendricks and Porter (1988). Here, truncation of the density typically will require the use of trimming or other data adjustment procedures to obtain an accurate representation of the density close to reserve prices. Subsequent work has explored some of the issues, but substantial problems remain in applying nonparametric techniques to standard auction data sets with differing number of bidders and substantial observed bidder heterogeneity.

The structural modeling literature on auctions also has been concerned with the more general question of whether it is possible to use bid data to discriminate between different private information specifications. Given the analysis above, such questions would seem to be relatively easy to resolve by matching observables to unobservables. For example, it seems at first glance plausible that a general AV model is unidentified because one only has N bids from which to infer the $N + 1$ unobservables – the N costs $c_1, c_2, \ldots, c_N$ and the general valuation v. Laffont and Vuong (1996) were the first to consider this question more formally and establish nonparametric identification results. They showed that for the same number of risk neutral bidders N, that any symmetric

AV model was observationally equivalent to some symmetric APV model. Moreover, they showed that while the symmetric APV and IPV models were nonparamertrically identified, the symmetric common values model was generally unidentified. Athey and Haile (2002) and others have examined the sensitivity of these results to different modeling assumptions and data sets. In particular, several authors have considered whether variation in the number of bidders can add additional identifying information.

8.1.3.2. Pure common value auctions In a pure common value auction, each bidder's private information, s_i, is an imperfect signal about the future value of the item. This additional source of uncertainty introduces another object of estimation – bidders' prior distribution on the value of the item, $f_v(v)$. To see how inferences are made about this prior and the density of private information given v, we consider a common values procurement auction.

In a pure common values procurement auction, all bidders have the same *ex post* cost c of performing a task. By assumption, each bidder j has an unbiased cost signal s_j of the cost of the project. This signal has marginal density $f(s_j \mid c)$ and conditional distribution function $F(s_j \mid c)$. In a PCV model, the bidders' private information signals are assumed conditionally independent and all agents are assumed to have the same prior $f_c(c)$ on the cost c.

In a Bayesian–Nash equilibrium, bidder j chooses b_j to solve the following expected profit maximization problem:

$$\max_{b_j} \Pi(b_j, s_j) = \int_{-\infty}^{\infty} (b_j - c)\big[1 - F\big(\beta^{-1}(b_j) \mid c\big)\big]^{N-1} h(c \mid s_j)\,\mathrm{d}c. \tag{106}$$

In this maximization problem, $b_j - c$ is bidder j's profit from winning, $[1 - F(\beta^{-1}(b_j) \mid c)]^{N-1}$ is bidder j's probability of winning given cost c, and $h(c \mid s_j)$ is the posterior density of c given the signal s_j. Bidder j's posterior density is

$$h(c \mid s_j) = \frac{f(s_j \mid c) f_c(c)}{\int_{-\infty}^{\infty} f(s_j \mid c) f_c(c)\,\mathrm{d}c}. \tag{107}$$

The symmetric Bayesian–Nash equilibrium bid function $\beta_c(s_j)$ is obtained from the first-order condition for the maximization problem. It satisfies the following differential equation

$$\beta_c'(s_j) - \beta_c(s_j) p(s_j) = q(s_j),$$

where

$$p(s_j) = \frac{\int_{-\infty}^{\infty} (N-1)[1 - F(s_j \mid c)]^{N-2} f^2(s_j \mid c) f_c(c)\,\mathrm{d}c}{\int_{-\infty}^{\infty} [1 - F(s_j \mid c)]^{N-1} f(s_j \mid c) f_c(c)\,\mathrm{d}c},$$

$$q(s_j) = -\frac{\int_{-\infty}^{\infty} c(N-1)[1 - F(s_j \mid c)]^{N-2} f^2(s_j \mid c) f_c(c)\,\mathrm{d}c}{\int_{-\infty}^{\infty} [1 - F(s_j \mid c)]^{N-1} f(s_j \mid c) f_c(c)\,\mathrm{d}c}.$$

This is an ordinary linear differential equation with solution

$$\beta_c(s_j) = \frac{1}{r(s_j)}\left[\int_{-\infty}^{s_j} r(u)q(u)\,\mathrm{d}u + k\right], \quad \text{with } r(\tau) = \exp\left(\int_{-\infty}^{\tau} p(u)\,\mathrm{d}u\right). \tag{108}$$

The constant k is determined by boundary conditions.

At first it might seem, following our discussion of the IPV model, that a researcher could use these integral equations as a basis for nonparametric estimation. Closer inspection of the differential equation reveals that for a given bid function, $\beta_C(s_j)$, there are a number of distribution functions $f(s_j \mid c)$ and $f_c(c)$ that could satisfy the above differential equation. This is in fact the nonidentification result of Laffont and Vuong (1996).

8.2. *Further issues*

These discussions of nonparametric identification show that identification can hinge delicately on any of several stochastic and economic assumptions. Indeed, there remain a great many combinations of auction formats and assumptions yet to be explored in the literature. For example, there are few general results on what can be identified with risk aversion. What results we do currently have suggest that much stronger identifying assumptions will be required when bidders are risk averse. [See Campo et al. (2003).]

It also is important to realize that most auction models in the theoretical and empirical literatures maintain that bidders' beliefs are symmetric. When bidders' beliefs differ for observable and unobservable reasons, auction models become much more challenging – both because it is more difficult to compute pure-strategy equilibrium bids and because there may be no pure-strategy equilibrium bids.

There also remain many institutional details that have yet to be fully explored in the nonparametric identification literature. For example, the presence of reserve prices can complicate both equilibrium bid functions and nonparametric estimation. These complications can destroy the identification of part or all of the relevant distributions of signals and common values. Another important assumption that may not hold in practice is the assumption that the number of bidders N is exogenous and known by the researcher. In many auctions, there appear to be few limitations on who can bid. One reason presumably why we do not see hundreds of bidders is because many are confident that their probability of winning is sufficiently low that this does not justify the expense of preparing and submitting a bid. Additionally, potential bidders could be deterred by the knowledge that other bidders are participating in the auction.

Despite all these limitations, nonparametric identification results and nonparametric estimation methods provide a useful reference for understanding what can be identified by imposing minimal economic rationality on observed bids. We now briefly consider what additional information can be gained by imposing parametric structure.

8.3. Parametric specifications for auction market equilibria

The previous subsection showed how structural econometric modelers have used first-order conditions from static auction models to estimate the primitives of alternative auction models. These first-order conditions had the form

$$b_j = \beta_j(v_j, N, F_j), \tag{109}$$

where v_j was a private valuation or signal, N is the number of potential or actual bidders, and $F_j(\cdot)$ was a joint distribution of private information and common uncertainties. From this equation we see that it is both bidder j's private valuation or signal v_j as well as bidder j's beliefs about other bidders' private information and common uncertainties, $F_j(\cdot)$, that affect observed bids. Under certain (identification) conditions, the system of bid functions can be inverted to recover the v_j

$$\begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix} = \begin{bmatrix} \beta_1^{-1}(B, N, F_1, \ldots, F_N) \\ \vdots \\ \beta_N^{-1}(B, N, F_1, \ldots, F_N) \end{bmatrix}$$

but only provided the bidder beliefs, F_j, are known or can be consistently estimated.

Because in a typical application F_j is unknown, it seems highly desirable that empiricists be as flexible as possible when estimating F_j. As we indicated repeatedly above, this desire raises a paradox: the cost of statistical flexibility may be economic flexibility. For example, to even begin to apply nonparametric techniques we must impose symmetry, $F_j = F_k$. Further, researchers typically do not have sufficient data to estimate general F's when N varies considerably across auctions or when there are many variables that enter the bid function (109). For this reason alone, many researchers have been willing to entertain parametric specifications for F. There are additional reasons to favor parametric specifications. One important one is that parametric specifications can identify economic quantities that are nonparametrically underidentified.

Some empirical researchers feel that as a matter of principle if something is not identified nonparametrically, one should never make parametric assumptions to identify it. Other researchers favor such restrictions if they lead to useful parameter estimates or counterfactuals. We hope it is clear by now that our position is that it is acceptable to make parametric assumptions as long as these assumptions are economically sensible and do not contradict the data. To appreciate the trade-offs that can arise in adding parametric structure, it is useful to see the trade-offs that Paarsch (1997) considered when developing a structural model of British Columbia government timber auctions. Paarsch's goal was to estimate a model of open-outcry timber auction within which he could ask whether the observed government reserve prices were revenue-maximizing. This is an ideal setting in which to use a structural model, because Paarsch seeks to perform counterfactual comparisons.

Several practical realities prevent Paarsch from employing the nonparametric estimation procedures discussed in the previous subsection. First, Paarsch has data on fewer

than 200 auctions. With less than 200 auctions, he has little hope of obtaining sensible estimates of a high dimensional conditional bid density. Second, there are at least five important observable dimensions along which the timber tracts differ. These differences include: the species composition of the tract, the amount of each species on the tract, the distance of the tract to local mills, and potential nonlinearities in harvesting costs. Third, the presence of a reserve price in the timber auctions introduces the same truncation problems present in Hendricks and Porter's (1988) descriptive study of offshore oil and gas lease auctions. Because Paarsch does not observe bids below observed reserve prices, he cannot estimate $F(\cdot)$ nonparametrically in these regions and thus cannot evaluate the revenue consequences of lowering reserve prices. Fourth, although his original sample consists of open-outcry and sealed-bid auctions, he chooses to focus exclusively on the open-outcry auctions. In open-outcry auctions, the dynamics of bidding can affect the observed sequence of bids.

Individually and collectively, these practical realities force Paarsch into a parametric specification of bidders' valuations and harvesting costs. Moreover, these realities also appear to force stark assumptions in order to obtain an estimable model. For instance, while Hendricks and Porter's discussion of oil and gas leases might suggest timber auctions have common value components, Paarsch rules out this possibility for at least two reasons. First, as a practical matter Paarsch's previous work showed that the private value auction framework provided as good or better explanation of winning bids as a pure common value auction framework. Second, English auctions in which bidders have common values are much more difficult to model. In an open-outcry English auction in which bidders' values are affiliated, bidders will revise their beliefs and bids according to the bidding history – not just their own private signal. For these and perhaps other reasons, Paarsch is led to adopt an independent private values framework.

In addition to adopting a private values framework, Paarsch also adopts a particular auction format to model the open-outcry structure of his timber auctions. Specifically, he assumes bid data are recorded via a "button" English auction. In an English button auction, all bidders begin the auction in plain view of one another with their fingers on a button. They listen to the auctioneer continuously call out increasing prices starting from the reserve price. A bidder exits the bidding (permanently) by removing their finger from the button. The last bidder depressing the button wins the auction at a price equal to the second-to-last bidder's value (or cost).

It is not hard to see why Paarsch makes this assumption. Because bidders' valuations are independent and private, bidders do not update their beliefs about valuations during the bidding. Moreover, their equilibrium strategy is to stay in the auction until the bidding reaches their valuation, at which point they drop out. (The winning bidder of course drops out when the second-to-last bidder does.) How does this equilibrium map into Equation (109)? Now, for all losing bidders

$$b_j = \beta(v_j, N, F_j) = v_j. \tag{110}$$

Thus, Paarsch's IPV assumption, when combined with the button English auction assumption, allows Paarsch to recover the valuations of all but the winning bidder from

the observed bids. There is little to do in terms of estimation. From any and all of the losing bids it would be possible to estimate a symmetric distribution of private values $F(\cdot)$ nonparametrically were it not for the fact that $F(\cdot)$ is conditioned on a large number of observables that vary across auctions.[28]

The automatic recovery of values from bids should sound familiar. This was exactly the solution in Example 10 where we drew a parallel between a perfect-information Bertrand model and an IPV model in which all bidders knew all costs. Here, Paarsch can recover exactly the valuation (here, profits) of the second-highest bidder. Paarsch also observes the third-highest valuation, and so on. Thus, if Paarsch were only interested in recovering valuations from bids, he could effectively dispense with the private information assumption altogether. To perform his reserve price counterfactuals, he could simply treat $F(\cdot)$ as a statistical construct that captures not private information of the bidders but simply unobserved (to him) reasons why bidders' profits would differ across tracts.

Other aspects of Paarsch's application have a bearing on how Paarsch estimates and interprets $F(\cdot)$ however. One of these is the match between the button auction model and the way the auctions were run and data collected. In a button auction, the last "bid" of a bidder is the price at which the bidder removes their finger from the button. In an open-outcry auction, the last observed bid is the last oral bid. For a variety of reasons, bidders may space their bids in an open out-cry auction, yielding the possibility of nonuniform jumps in bids. If this is the case, it is unclear how one should interpret the last bids in Paarsch's data.

There are other features of the timber auctions that affect the empirical model. To appreciate these features, it is useful to go into the details of his application. Paarsch decomposes each bidder j's valuation v_{ij} into an average revenue per tree on the tract, r_i, and average harvesting costs, c_{ij}. That is, $v_{ij} = r_i - c_{ij}$. The absence of a bidder j subscript on revenues, and the lack of any common value component in the auction, immediately implies that revenues are known to all bidders. In addition, Paarsch assumes that he observes the revenues bidders observe. He calculates these revenues as a sum of species prices times the amount of each species the government estimates is on each tract. Thus, when it comes to distinguishing between private information or unobserved heterogeneity models, it is the individual differences in harvesting costs that are important.

A key novelty of Paarsch's paper is that he models the difference between the potential number of bidders in an auction and the number who end up bidding. To see why this distinction is important, notice that the reserve prices in Paarsch's auctions truncate not only the distribution of observed bids, but lead to a difference between the potential number, $\mathbf{N_i}$, and actual number, N_i, of bidders. To model this difference, Paarsch makes parametric assumptions about the distribution of bidders' harvesting costs and how they

[28] There is additional information in the condition that the winning bidder's valuation exceeds the winning bid. Paarsch could presumably use this inequality to improve the precision of estimates of $F(\cdot)$.

vary across auctions. Specifically, he introduces two types of bidder heterogeneity. In the leading case, he models a bidder's average cost c_{ij} as being drawn from a Weibull density

$$c_{ij} \sim F(c \mid \delta) = 1 - \exp\bigl(-\delta_1 (c - c_{\min})^{\delta_2}\bigr), \quad c \in [c_{\min}, \infty],$$

where the δ's are unknown parameters that affect the heterogeneity of costs and $c_{\min}$ is a lower bound equal to

$$c_{\min} = \gamma_0 \frac{1}{q} + \gamma_2 q + \gamma_3 q^2 + \gamma_4 d.$$

Here, q measures the size of the tract and d is the distance to the closest timber mill. The δ and γ parameters help capture reasons why the distribution of average harvesting costs might vary across auctions. Adequately capturing such variation would be crucial to accurate counterfactual calculations. In a second specification, Paarsch considers the consequences of assuming fixed harvesting costs, γ_0, are random.

By modeling costs parametrically, Paarsch can use maximum likelihood to estimate the unknown costs of bidders. These costs are critical to his optimal reserve price calculations. One major problem remains however – how to account for the fact that while he observes the number of bidders, N_i, he does not observe the number of potential bidders, $\mathbf{N_i}$.

To appreciate this problem, consider timber auctions that have at least 2 bids, i.e., $N_i \geqslant 2$. There will be a difference between N_i and $\mathbf{N_i}$ when potential bidders with extremely high harvesting costs find it unprofitable to bid above the reserve price. The likelihood of observing the order statistic data $c_{[2:\mathbf{N_i}]}, c_{[3:\mathbf{N_i}]}, \ldots, c_{[N_i:\mathbf{N_i}]}$ and N_i is

$$L(\gamma, \theta \mid C_i, N_i) = \binom{\mathbf{N_i}}{N_i} \bigl[1 - F(c^*)\bigr]^{\mathbf{N_i} - N_i} F(c^*)^{N_i} \times N_i! \frac{F(c_{[2:\mathbf{N_i}]})}{F(c^*)^{N_i}} \prod_{j=2}^{N_i} f(c_{[j:\mathbf{N_i}]}). \tag{111}$$

The first portion of this likelihood function (before the $\times$) is the (binomial) probability of observing $\mathbf{N_1} - N_i$ cost draws below the cost c^*, where c^* is the cost that would result in profit of zero at the reserve price. The second portion is the density of observable average harvesting costs given N_i and that the unobserved lowest cost satisfies $c_{[1:N]} < c_{[2:N]}$. The problem with trying to estimate δ and γ with this likelihood function is that Paarsch cannot compute the likelihood function unless he knows the number of potential bidders $\mathbf{N_i}$ for each auction. Because he does not know $\mathbf{N_i}$, he could treat each $\mathbf{N_i}$ as a parameter to be estimated. This, however, amounts to introducing a new parameter for each auction. As is recognized in the econometrics literature, the introduction of so many parameters will make the maximum likelihood estimator inconsistent. Absent a solution then to this problem, Paarsch, and many later auction researchers, are stuck.

A main contribution of Paarsch's paper is to show that a conditional likelihood function approach [Andersen (1970)] can be used to obtain consistent estimates of δ and γ. The conditional likelihood approach works as follows. Let $f(C_i, N_i \mid \mathbf{N_i})$ be the joint density of observed costs and bidders conditional on the unobserved potential number of bidders in auction i. According to the conditional maximum likelihood approach, if this density can be factored into two pieces of the form

$$f(C_i, N_i \mid \mathbf{N_i}, \delta, \gamma) = g(N_i \mid \mathbf{N_i}, \delta, \gamma) \times h(C_i \mid N_i, \delta, \gamma),$$

then one can obtain consistent estimates of δ and γ by maximizing the conditional likelihood function $h(C_i \mid N_i, \delta, \gamma)$. Paarsch's contribution is to show that for this specific IPV auction, the likelihood function (111) has this form, with N_i serving as a sufficient statistic for the unknown potential number of entrants.

We now are in a position to return to the point on which we began this example. While the costs of parametric assumptions in many applications are self-evident, the benefits are sometimes less clear. One important benefit of parametric structure is that it may allow the researcher to identify a quantity of interest. In Paarsch's case, the realities of timber auctions necessitated several strong modeling and parametric assumptions, such as private values and an English button format. On the other hand, the resulting model did overcome a significant handicap, which is that the number of potential bidders is rarely known.

Whether this benefit justifies the starkness of the assumptions, has to be viewed from at least three vantages. First, is the benefit practically useful? The answer here appears to be a resounding yes. Without it Paarsch could not estimate his model and perform the counterfactual optimal reserve price calculations. Second, does the parametric structure deliver the end result? In Paarsch's case, the answer is unclear. Finally, does the additional structure adequately capture the economics of the agents' behavior, particularly when it comes to the counterfactuals? To answer this question, Paarsch tries to convince readers by reporting alternative models and estimates.

8.4. Why estimate a structural auction model?

Previously, we asserted that researchers should not attempt a structural model without a convincing explanation of how its benefits will outweigh potentially restrictive and untestable assumptions. This advice seems particularly relevant when considering how to model auction bid data.

The main benefit of a structural auction model would seem to be that it allows the researcher to estimate the distribution of bidders' valuations (or similar objects). Such estimates can in principle be used to evaluate an auction's efficiency or how changes in the rules would affect the seller's revenues.

In actual applications, however, these benefits are only achieved at the cost of restrictions on bidders' information. In particular, the vast majority of structural auction models either exclusively estimate independent private values or pure common values models. The reasons for this specialization are not too hard to find – more realistic

affiliated models are analytical and computationally intractable.[29] Restrictions on the distribution of bidders' information naturally limit the applicability of the estimated model. For example, it makes little sense to estimate an IPV model and then use those estimates to model what would happen if there was a common value.

Even if we are willing to accept the independent private values or pure common values assumptions, there are other factors that can affect the value of structural estimates. Consider what we learn from estimating an IPV model. The best one can hope for is to be able to recover a precise nonparametric estimate of the distribution of bidder valuations $F(v_j)$ above for valuations that would lead to bids above the reserve price. But what is the value of knowing $F(v_j)$? We believe that the answer is that there is little or no value unless we can somehow say that $F(v_j)$ is applicable to past auctions or future auctions. For example, we could imagine estimating $F(v_j)$ during a period in which there was no collusion among bidders and then trying to use the estimated density to compare bids (valuations) when bidders were perhaps colluding. Alternatively, like Paarsch (1997), one could perform counterfactuals that involve changing some aspect of the auction like the reserve price.

The key question is: How does one know that the estimated valuation distribution is relevant to other auctions? Our position is that to be convincing, the structural modeler has to have a convincing explanation for when $F(v_j)$ is likely or unlikely to change from auction to auction. To take Paarsch's (1997) timber auction model as an example, we might ask: When would his estimates be relevant for a timber auction in another Canadian province? To answer this question, we ultimately would need to understand how timber auctions are different. This is not a question that auction theory itself can answer directly. Instead, the answer likely lies in the specifics of what is being auctioned and how it is auctioned. Thus, we see that economic theory often can only go so far in answering specification issues. In the end, the econometrician will have to pick and justify conditioning variables. Ideally, these choices will be made with the aid of economics, but in practice it is knowledge of the industry, institutions and data that will likely make the analysis convincing.

Suppose we can accept the assumptions of a structural auction model, what can we do with the resulting estimates? Structural auction models can in principle facilitate useful counterfactual experiments. Paarsch's (1997) evaluation of optimal reserve prices is one example. Other researchers have used structural models to evaluate alternative winning bid rules. One area where structural auction models have yet to make much headway is in diagnosing bidder collusion. Here there are two problems. First, economists do not have particularly good models of how colluding bidders behave. Indeed, the modeler often is confronted with the paradox: rationality suggests that colluding bidders will scramble their bids so as to make detection extremely difficult. To date, most of what

[29] Although there have been some attempts to compare private values and common values models, these tests invariably rest heavily on functional form and other assumptions. In the end, little progress has been made using structural models to decide the appropriateness of different information structures.

structural models have contributed to the detecting collusion literature are benchmark noncooperative models of bids. IO economists have used these models to look for suspect bid clustering, skewing or correlation.

There are additional practical issues that limit the usefulness of structural auction models. One set pertains to dynamic considerations. In many applications, the data come from repeated auctions where the same bidders bid against one another. This repetition raises two issues. First, in repeated auctions, bidders' subsequent valuations may be influenced by the number of units they have won in the past. In this case, symmetric information models no longer make sense. Second, in repeated auctions bidders likely will internalize information and strategic externalities that their bids today may have for bids tomorrow.

8.5. *Extensions of basic auctions models*

Recent research has addressed many of the limitations associated with the auction frameworks described in Table 1. It is well beyond the scope of this chapter to even begin to survey this literature. Interested readers should consult Hendricks and Porter (in press).

There are some developments that fit in with our earlier discussions that are worth noting briefly. Laffont, Ossard and Vuong (1995), for example, extended the IPV paradigm to allow for both observable and unobservable heterogeneity across auctions. Although their estimation procedure assumes a parametric model for the distribution of private valuations, they devise a clever estimation technique based on simulated nonlinear least-squares that does not require them to compute the equilibrium bid functions. Instead their technique simulates the expected value of the winning bid for an arbitrary distribution of private values and a potentially binding reserve price. They also treat the number of potential bidders as a random variable.

Haile and Tamer (2003) explore the empirical implications of English auctions. The button English auctions we considered earlier are a special type of English auction. In open-outcry English auctions, bidders can bid whenever they are willing to best the outstanding bid (plus any minimum bid increment). Exactly what order and when bidders will bid is something left to the auction's format and strategic considerations. In general, the dynamics of English auctions are extremely difficult to analyze. Rather than try and detail the dynamics of the bidding, Haile and Tamer take a minimalist approach by using potentially weak restrictions on players' bids. Specifically, Haile and Tamer maintain that observed bids need only satisfy the following two restrictions: (1) bidders do not bid more than they are willing to pay; and, (2) bidders do not allow an opponent to win at a price they are willing to beat. Using these assumptions, they derive bounds on the distribution of valuations and bids above reserve prices. These bounds become exact for a button auction and are weak bounds for other English auctions.

There has also been recent empirical research on multi-unit auctions. Virtually all wholesale electricity markets operating around the world run daily multi-unit auctions to determine which generation facilities are able to supply energy. Each day suppliers

submit nondecreasing step functions expressing their willingness each hour to supply electricity for the next 24 hours. The system operator then computes the least cost way to meet demand in each hour based on these bids. Wolak (2000) develops a model of expected profit-maximizing bidding behavior in such markets. Wolak (2003) uses this model to estimate bidder cost functions. He shows that because of the richness of the bid functions that market participants submit, the only assumption required to recover these cost function estimates is expected profit maximizing bidding behavior. An important difference between these multiple-good auctions and single good auctions is that in a multi-unit auction, suppliers compete over how many units they sell. Consequently, residual demand (market demand less the willingness to supply functions of all other market participants) is observable *ex post*, and this provides the information necessary to identify the supplier's underlying marginal cost function.

As should be clear from this brief discussion, significant progress has been made in deriving flexible modeling frameworks which allow empirical IO researchers to recover information about the distribution of private information in auction models under minimal assumptions.

9. Games with incomplete information: Principal-agent contracting models

Recently, IO economists have begun to develop structural econometric models of regulator and regulated firm interactions. These empirical models are more ambitious than the auction or oligopoly models discussed in the previous sections. Similar to oligopoly models but unlike auction models, these models seek to estimate production and demand functions. Similar to auction models but unlike most oligopoly models, these models seek to account for the impact of asymmetric information on agents' strategic interactions. These ambitious modeling goals usually require the researcher to rely on stronger parametric and distributional assumptions to identify and estimate economic primitives. The main goal of this section is to discuss why models of regulatory interactions require this structure.

As in auctions, private information plays a critical role in regulatory proceedings. IO economists have recently used principal-agent contracting models to characterize regulatory proceedings in which regulators set the prices (or "rates") regulated firms (or "utilities") charge. A key insight from these models is that when a utility has superior information about the underlying economic environment, it can exploit that information to earn greater profits than it would if the regulator were equally informed. This paradigm for studying regulator–utility interactions has received such widespread acceptance among IO economists that Laffont and Tirole (1993) have coined the phrase "new regulatory economics".

One of the most important economic primitives in these contracting models is the economist's specification of the regulated firm's private information. The two main types of private information a utility can have are private information about its production process or its demand. The regulated firm has no incentive to reveal this private

information to the regulator because the regulator would use this information against them in the rate-setting process. The regulator in turn is aware that the firm has private information, and takes this into account in setting rates. Economic theorists model this interaction by computing optimal "second-best" solutions to a revelation game. In this game, the regulator announces a price schedule and a transfer payment that are functions of the firm's reported private information.

A critical constraint on the firm is that it must serve all demand consistent with the private information it reports. (Its private information determines the price granted by the regulator.) Under an optimal "second-best" solution, the price schedule chosen by the regulator maximizes a social welfare function subject to the constraints that: (1) the firm finds it profit maximizing to report its true private information to the regulator; and, (2) the firm expects to earn profits sufficient to keep it from exiting the industry. Although these theoretical models are stylized static depictions of regulatory interactions, they do capture important features of the asymmetric information problem faced by actual regulators. Important examples of this work include Baron and Myerson (1982), Baron and Besanko (1984, 1987), Besanko (1984) and Laffont and Tirole (1986).

Historically, empirical IO economists have largely ignored the impact of regulated firms' private information on both regulated firm and regulator behavior. Instead, empirical IO economists have estimated conventional cost and demand functions using standard cost functions, factor demand equations and product demand models. Christensen and Greene's (1976) study of electric utility costs is a classic example of regulated firm cost function estimation. Evans and Heckman (1984) provide a more recent example of cost function estimation applied to the AT&T divestiture decision. In virtually all cost function studies, statistical tests of cost-minimizing behavior are rejected.

The rejection of cost-minimizing behavior is not too surprising if one recognizes the presence of private information. A regulated firm with private information need not find it profit-maximizing to minimize costs if it can distort its behavior to obtain better prices from the regulator. Given these incentives, estimation procedures that assume cost minimization behavior will yield inconsistent estimates of the underlying economic primitives.

The remainder of this section follows the format of the previous section. First, we describe the data a researcher has in a typical application. We then develop a simple model that illustrates what economic primitives can be recovered from these data. After considering nonparametric identification, we discuss the practical limitations of nonparametric identification results. This then leads us to describe how parametric assumptions can be used to identify economic primitives. We illustrate this discussion using Wolak's 1994 study of Class A California Water Utilities. We close with a short discussion of subsequent related empirical work.

9.1. Observables and unobservables

Empirical research on regulated industries benefits from regulatory proceedings that make rich cost and revenue data publically available. On the cost side, for example,

regulated utilities typically must report detailed data on inputs, X, and input prices, p_X. Inputs consist of information on the firm's capital (K), labor (L), energy (E) and materials (M) choices associated with an observed output Q. Additionally, the researcher also will have information on input prices, $p_X = (p_K, p_L, p_E, p_M)'$. Using these data, a researcher can construct an estimate of the total cost, C, of producing the observed output level Q. In terms of the above notation

$$C = p_K K + p_L L + p_E E + p_M M. \tag{112}$$

On the output (or revenue) side, firms provide both retrospective and prospective quantity and revenue data. The prospective quantity data reflect the reality that the regulator must set prices before either it or the firm know what demand will be. When setting price, the regulator attempts to balance two competing goals: (1) it must allow the firm to recover all "prudently" incurred costs; and, (2) it must provide strong incentives for the firm to produce in an efficient manner. To model the prospective nature of the regulator's pricing decisions, it is imagined that demand equals $D(p_Q, Z, \epsilon_Q) = Q$, where p_Q is an output price set by the regulator, Z is a vector of observable variables assumed to shift demand and ϵ_Q is an unobserved demand shifter.

Regulatory models differ according to whether ϵ_Q is known to the firm (i.e., is private information) or is unknown to the firm before the firm reports to the regulator. In what follows, we only explore models in which the firm has private information about its production function. Thus, ϵ_Q here does not reflect private information. The econometrician of course never observes ϵ_Q.

Given these cost and output data, all an empirical researcher can do is consistently estimate the joint density of regulated prices, firm outputs, firm input choices, and total costs – conditional on input prices (p_X) and any demand shifters (Z); i.e., the researcher can estimate $h(p_Q, Q, X, C \mid p_X, Z)$. Input prices and the demand observables are used as conditioning variables because firms are thought to be unable to impact input prices or factors that influence demand. Thus, these vectors Z and p_X are usually assumed to be distributed independently of all of the unobservables in the econometric model.

To obtain a consistent estimate of the firm's production process, the researcher must be very specific about how the utility's private information interacts with the regulatory process. In what follows, we explore models in which the regulated firm has private information about its production process. We restrict our attention to private information on the production side in keeping with Wolak's (1994) empirical model. Specifically, we model the firm's private information as a single parameter that enters the firm's production function $Q = f(K, L, E, M, \theta)$. The firm knows θ from the start and all the regulator knows at the start is the density of θ, $f_\theta(\theta)$, where $\theta \in [\theta_l, \theta_h]$. Absent further assumptions on the distributions of θ and ϵ_Q, and specific functional forms for $D(p_Q, Z, \epsilon_Q)$ and $f(K, L, E, M, \theta)$, little or nothing can be deduced about these underlying economic primitives from $h(p_Q, Q, X, C \mid p_X, Z)$. This is because the firm's input choices will depend in an unknown way on θ, which implies that total cost, C,

does as well. Additionally, because the firm must by law satisfy all demand at the regulated price, the firm's output will depend on the realization of ϵ_Q, the unobservable demand shifter. This implies that the firm's input choices and total cost will also be a functions of the realization of ϵ_Q. Consequently, without functional form restrictions on the demand and production functions, or assumptions about the forms of the distributions of θ and ϵ_Q, the researcher will be unable to identify demand and cost functions from $h(p_Q, Q, X, C \mid p_X, Z)$.

These observations lead us to consider the types of functional form and distributional assumptions that can lead to identification. We will see that nonparametric identification of the distribution of private information, as in independent private values auction models, hinges on a monotonicity condition. We show that strong economic or statistical assumptions are required to guarantee monotonicity. We then discuss parametric models. These models rely on functional form and distributional assumptions to identify the underlying economic and information primitives.

9.2. *Economic models of regulator–utility interactions*

Baron (1989) provides a useful theoretical model for thinking about empirical models of regulator and utility interactions. He assumes $C(q, \theta) = \theta q + K$ where θ is the firm's private marginal cost of producing output q. No explicit economic rationale is provided for the cost function. In particular, there is no reason to believe that the firm produces its output at minimum cost for any value of θ. In this sense, we can think of $C(q, \theta)$ as a behavioral cost function; it gives the cost of producing output q given θ.[30] Additionally, Baron assumes $D(p)$ represents the quantity demanded at the regulated price p. Thus, in his model there is no demand uncertainty.

In Baron's model, the regulator fixes a price schedule, $p(\theta)$, and a monthly (or annual) fixed fee schedule, $T(\theta)$, that give prices and fixed fees as a function of the firm's announced marginal cost θ. Given the price and fixed fee schedules, the firm announces a marginal cost, $\hat{\theta}$, to maximize its profits

$$\pi(\hat{\theta}; \theta) = p(\hat{\theta}) D\big(p(\hat{\theta})\big) + T(\hat{\theta}) - \theta D\big(p(\hat{\theta})\big) - K. \tag{113}$$

There are two constraints imposed on the regulator's price and fee optimization problem. The first is a truth-telling or incentive compatibility constraint. This constraint requires that a firm of type θ will report its true type. In other words, a truthful report must yield the firm profits that are greater than or equal to profits it could obtain through any other feasible report in the support of θ. Mathematically, this implies:

$$\pi(\theta) \equiv \pi(\theta; \theta) \geqslant \pi(\hat{\theta}, \theta), \quad \forall \hat{\theta} \in [\theta_l, \theta_h] \text{ and } \forall \theta \in [\theta_l, \theta_h]. \tag{114}$$

[30] By behavioral cost function we mean only that the firm behaves according to a consistent set of rules that yield this stable relationship between costs and q for a given value of θ. One possible set of behavioral rules is to minimize total production costs, but this is not necessary because, as discussed above, the firm may have little incentive to produce its output according to minimum cost.

As Baron notes, these constraints are global. That is, they must be satisfied for each θ and all feasible reports $\hat{\theta}$.

The second constraint is called the participation constraint or individual rationality constraint. It states that regardless of the firm's true value of θ, it must receive more than its outside option. Here this means that the firm must earn nonnegative profits. Mathematically,

$$\pi(\theta) \geqslant 0, \quad \forall \theta \in [\theta_l, \theta_h]. \tag{115}$$

Because it is extremely complicated to impose the global truth-telling constraint on the regulator's optimization problem, theorists typically make assumptions about economic primitives so that satisfaction of local truth-telling implies satisfaction of global truth-telling. These assumptions are analogous to those in auction models that make the bid functions monotone in the bidders' valuations.

Baron (1989, pp. 1366–1367) shows that the local truth-telling constraint for this problem is the following differential equation in θ:

$$\frac{\mathrm{d}\pi(\theta)}{\mathrm{d}\theta} = -C_\theta\big(D\big(p(\theta)\big), \theta\big). \tag{116}$$

This equation tells how profits must increase as a function of θ in order to induce local truth telling. In words, for small deviations from truthful reporting, the firm experiences a decline in profits. This condition can easily be checked for the assumed cost function, as $C_\theta = q > 0$ for all θ.

As Baron notes, Equation (116) can be integrated to produce an expression for the firm's profit

$$\pi(\theta) = \int_\theta^{\theta_h} C_\theta\big(D\big(p(x)\big), x\big)\,\mathrm{d}x + \pi(\theta_h). \tag{117}$$

This equation implies that the participant constraint can be simplified to

$$\pi(\theta_h) \geqslant 0, \tag{118}$$

which means that the least efficient firm, as parameterized by θ, must earn nonnegative profits. Using the definition of $\pi(\theta)$ in Equation (114), we can re-write Equation (113) as

$$\pi(\theta) = p(\theta)D\big(p(\theta)\big) + T(\theta) - \theta D\big(p(\theta)\big) - K. \tag{119}$$

Deriving the optimal price and fixed fee functions requires specifying the regulator's objective function. The general objective function considered for the regulator is a weighted sum of consumer and producer surplus. Because both consumer and producer surplus will depend on the firm's actions, which depend on the unobserved θ, the regulator must use its knowledge of $f(\theta)$ to compute an expected surplus function

$$W = \int_\theta^{\theta_h} \left[\int_{p(\theta)}^{\infty} D(x)\,\mathrm{d}x - T(\theta) + \alpha\pi(\theta) \right] f(\theta)\,\mathrm{d}\theta, \tag{120}$$

where α is the relative weight given to the firm's profits in the regulator's objective function. The regulator is assumed to choose the price and fixed fee schedules to maximize (120) subject to (115), (116), and (119) using calculus of variations techniques.

Baron (1989) shows that the optimal price schedule takes the form

$$p(\theta) = \theta + (1-\alpha)\frac{F(\theta)}{f(\theta)}, \tag{121}$$

which looks very similar to Equation (99) in the independent private values auction case, apart from the parameter α. Baron shows that a sufficient condition for satisfaction of the local truth-telling constraint to imply satisfaction of the global truth-telling constraint is that $p(\theta)$ is nondecreasing in θ. This equation shows that monotonicity of the price function imposes restrictions on the distribution of θ. Specifically, if $F(\theta)/f(\theta)$ is nondecreasing in θ, then $p(\theta)$ is nondecreasing in θ.

If the value of α is known to the econometrician and firms face the same cost function and nonstochastic demand function, then it is possible to recover a consistent estimate of the density of θ, $f(\theta)$, from prices. Such an exercise would follow the "change-of-variables" logic applied to the first-order condition in sealed-bid IPV auction models. It is important to emphasize all of the assumptions necessary for this identification result. Besides assuming firms have the same cost function and density of private information, we have assumed the demand function is the same across all observations. In other words, $D(p)$ cannot vary across observations and there are no unobservable ϵ_Q or observable demand shifters. Equation (121) also depends crucially on the functional form of $C(q,\theta)$. Without the constant marginal cost assumption, the regulator's optimal price schedule will depend on the demand function $D(p)$.[31]

Although this nonparametric identification result may at first seem appealing, it should not give much comfort to regulatory economics researchers for three reasons. First, it is difficult to imagine circumstances where the researcher will know the value of α. Second, the underlying cost function tells the researcher nothing about the technology of production. As noted earlier, $C(q,\theta)$ simply characterizes the relationship between production costs, q, and θ. The researcher cannot say anything about the returns to scale in production, the elasticity of substitution between inputs or the extent to which the regulatory process results in deviations from minimum cost production. Moreover, the manner in which θ enters the cost function is extremely restrictive. Third, nonparametric identification rests on unrealistic assumptions about the extent of observed and unobserved heterogeneity in the production process and demand. Specifically, in this model the only reason market prices differ across observations is because of different realizations of θ. It is difficult to imagine a sample of regulator–utility interactions with no observed or unobserved heterogeneity in the production and demand functions.

[31] See the discussion of Wolak (1994) below. In addition to recovering the density $f(\cdot)$ nonparametrically, it is possible to recover a consistent estimate of K from information on the regulated quantity and total production cost. Also, if the researcher is willing to assume $D(p)$ is the same for all observations in the sample, then the set of observed (p_Q, Q) pairs will nonparametrically trace out the demand curve $D(p)$.

Some of these shortcomings can be overcome by explicitly specifying an underlying production function and how it depends on θ. The firm's observed cost function can then be derived from the assumption of expected profit-maximizing behavior subject to the constraints imposed on firm behavior by the regulatory process. Both observed and unobserved heterogeneity can also be allowed in the production function and the demand function facing the regulated firm. However, there is a cost of being more general – nonparametric identification is lost, just as it is in the case of auction models. As we now show, however, by clever choices of functional forms and distributional assumptions, the researcher can estimate a rich set of underlying economic primitives.

9.3. Estimating productions functions accounting for private information

Wolak (1994) derives and implements a procedure to recover a consistent estimate of a regulated firm's production technology taking into account the impact of private information on regulator–utility interactions. As noted above, this task requires the imposition of parametric and distributional assumptions. These assumptions allow Wolak to identify the underlying economic primitives from the joint density of the regulated price, the firm's output, input choices and total cost, conditional on the vectors of input prices and demand shifters, $h(p_Q, Q, X, C \mid p_X, Z)$. As we have said repeatedly, there is no single "right" way to make these assumptions. It is presumably economics and the specific features of a market and regulatory environment that can help the researcher defend the assumptions necessary to obtain identification.

Wolak models the behavior of a sample of Class A California Water utilities using annual data on utility outputs, production costs, input quantities and prices, several demand shifters, and output prices. He has panel data from 1980 to 1986. Class A utilities distribute water and provide services to large cities in California. Consistent with our earlier discussion, the California Public Utilities Commission (CPUC) sets the retail price of water for these utilities on a prospective basis.

As Wolak (1994) notes, the water supply industry was chosen, as opposed to other regulated industries, such as telecommunications, or electricity, for two major reasons. First, the structure of production in water delivery is extremely simple relative to producing electricity or providing telecommunications services. Second, the assumption of a single homogenous product is likely to be far less objectionable than would be the case for either telecommunications or electricity. These reasons help Wolak to simplify his econometric model.

As with any structural econometric modeling exercise, it is important to have a clear idea of what economic magnitudes can be recovered from a structural model. Wolak would first like to obtain a consistent estimate of the underlying production function. To do this, he explicitly models the impact of the utility's private information on production. Instead of estimating the production function directly, Wolak derives the utility's cost function under the assumption of expected profit-maximizing behavior. He then estimates the production function parameters from the cost function. A useful by-product of this approach is an estimate of the distribution of private information. A second goal

of the structural model is to obtain an estimate of how much firm output is distorted from minimum cost production due to the presence of private information. A third goal is to test the relative performance of the asymmetric information model versus the conventional symmetric information model of the regulator–utility interaction.

To evaluate the relative performance of the asymmetric information model, the paper posits a second behavioral model of regulator–utility interaction for the same set of underlying economic primitives. In this model, the utility initially possesses private information. Through information gathering, however, the regulator is able to completely learn this parameter. Consequently, the regulator can impose what Wolak calls the symmetric (or full) information regulatory outcome. Unfortunately, the econometrician is unable to observe this private information parameter and so must take it into account.

Wolak does this when specifying and estimating the behavioral cost functions. The asymmetric information model assumes that the utility possesses private information, but the regulator is unable to completely learn this private information through its information gathering efforts. However, the regulator does learn the distribution of this private information for each utility, and regulates using this incomplete information optimally. The regulator is assumed to impose a version of the asymmetric information optimal "second-best" regulatory outcome described in the previous section. In this case, the econometrician also is unable to observe the utility's private information (or even its distribution), but must account for this assumed utility–regulator interaction when estimating the parameters of the utility's production function.

Wolak assumes the production function for water delivery for utility i is

$$Q_i = f\left(K_i, L_i^*, E_i, \epsilon_i^Q \mid \beta\right), \tag{122}$$

where K_i denotes capital (physical plant and water sources), L_i^* labor, and E_i electricity. The parameter β is a vector describing the technical coefficients of production. It is known to both the regulator and utility, but is unknown to the econometrician. The variable ϵ_i^Q is a stochastic disturbance to the ith utility's production process that is realized after the utility makes its capital stock selection, but before it produces. The utility knows the distribution of ϵ_i^Q, which is independently and identically distributed over time and across utilities. Allowing for this source of unobservable heterogeneity in the production function increases the realism of the model because there are a number of factors that are unknown to the firm at the time it chooses the configuration and capacity of its water distribution network. (A utility's distribution network is a major component of its capital stock.)

To account for all these forms of unobserved heterogeneity, Wolak must make parametric and distributional assumptions to identify the underlying economic primitives from $h(p_Q, Q, X, C \mid p_X, Z)$. Without these assumptions, it is impossible to proceed. Once again, this illustrates our point that it is specific parametric economic and statistical assumptions that allow us to go from the statistical joint distribution of the data, $h(p_Q, Q, X, C \mid p_X, Z)$, to statements about the production technologies, market demand and information primitives.

The source of the utility's private information is the efficiency of its labor input. To this end, Wolak makes the distinction between, L_i^*, the amount of labor actually used in the production process, and L_i, the observed physical quantity of labor input which is implied by the utility's total labor costs. These two magnitudes are related by the equation $L_i^* = L_i/\mathrm{d}(\theta_i)$, where $\mathrm{d}(\theta)$ is a known increasing function and θ_i is interpreted as utility i's labor inefficiency parameter. (Higher values of θ_i imply more inefficiency.) The econometrician and regulator observe the utility using the quantity of labor L_i, but the actual amount of "standardized" labor available in the production process is L_i^*. This specification is based on the fact that labor costs are a major component of total maintenance expenditures, and system maintenance is a major determinant of water system efficiency. Thus, while the regulator can observe how much labor is employed at the utility, L_i, it does not know the productivity of this labor. The utility's *ex post* observed costs have the form $w_i L_i + r_i K_i + pe_i E_i$, where w_i is the wage rate, r_i is the price of capital, and pe_i is the price of electricity. Note that the utility pays for observed labor, L_i.

From the viewpoint of the econometrician, θ_i is an unobservable random variable that determines the productivity of labor. In this sense, it is comparable to other unobservables, such as ϵ_i^Q. What is special about θ_i as an unobservable is that it may also be unobserved by the regulator. This is the case of Model A, the asymmetric information model. There, the regulator only knows the distribution of θ_i, $F(\theta)$, and thus can only condition its decisions on that information – much like what happens in an auction. By contrast, in the symmetric information model (Model S), the θ_i plays the role of unobserved heterogeneity because the regulator and firm observe it but the econometrician does not.

For both Model S and Model A, the utility chooses its input mix to maximize expected profits given its private information. Each utility faces the demand function $Q^D = Q_i(p_i)\epsilon_i^D$ for its product, where ϵ_i^D is a positive, mean one stochastic shock to demand. This shock is assumed independently and identically distributed across time and utilities. Once p is set, the demand shock is realized; the utility then produces output to satisfy demand (which in both models is known both to the regulator and the utility).

Because the utility's price and capital stock are set before the utility produces each period, the utility's desire to maximize expected profits will lead it to minimize total operating costs under both Models A and S for a fixed level of output and capital stock. Thus, for each model, Wolak can compute a conditional variable cost function $\mathrm{CVC}(pe, w, \theta, K, Q, \epsilon^Q, \eta_L, \eta_E \mid \beta)$, where η_L and η_E are mean one optimization errors. Wolak introduces these errors to allow for the fact that the first-order conditions for L and E do not hold exactly. Note that the utility's private information, θ, enters into the conditional variable cost function.

Using this expression for variable costs, utility i's total observed costs equal:

$$\mathrm{TC} = \mathrm{CVC}(pe, w, \theta, K, Q, \epsilon_q, \eta_L, \eta_E \mid \beta) + r_i K_i. \tag{123}$$

As noted earlier, the firm's capital stock serves two roles: (1) it reduces the total cost of serving demand; and, (2) it signals to the regulator the firm's true productive efficiency. This tension between increasing profits by choosing the minimum total cost level of capital and increasing the size of the capital stock in an effort to be rewarded by the regulator with a higher output price, leads to distortions from least-cost production by the firm.

9.3.1. Symmetric information model

In the symmetric information model, the regulator observes each utility's true θ and sets the monthly fixed fee (F) and per unit price (p) to maximize expected consumer surplus subject to the constraint that the utility's expected profits (with respect to the distributions of ϵ^Q and ϵ^D) equal zero. This implies that the regulator will solve for the p, T, and K which maximize expected consumer surplus for the utility's consumers.

Let $S_i(p) = E_D(\epsilon_i^D)\int_p^\infty Q_i(s)\,ds$ denote expected consumer surplus for the ith utility, where $E_D(.)$ denotes the expectation with respect to the distribution of ϵ^D. In terms of our notation, the regulator solves:

$$\begin{aligned}
&\max_{p,T,K} S_i\big[p(\theta_i)\big] - T(\theta_i)\\
&\text{subject to}\quad E_{QD}\big(\pi(\theta_i)\big) = E_{Qd}\big[p(\theta_i)Q\big[p(\theta_i)\big]\epsilon_i^D + T(\theta_i)\\
&\qquad\qquad - \mathrm{CVC}\big(pe, w, \theta_i, K(\theta_i), Q(\theta_i)\epsilon_i^D, \epsilon_i^Q, \eta_i \mid \beta\big)\big]\\
&\qquad\qquad - r_i K(\theta_i) = 0, \qquad (124)
\end{aligned}$$

where $E_{Qd}(.)$ is the expectation with respect to the distribution of both ϵ^Q and ϵ^D and $\eta_i = (\eta_i^L, \eta_i^E)'$ is the vector of optimization errors from the conditional variable cost function optimization problem. The first-order conditions for the regulator's problem imply:

$$p_i = \frac{\partial E_{QD}[\mathrm{CVC}(pe, w, \theta_i, K(\theta_i), Q(\theta_i), \epsilon_i^D, \epsilon_i^Q, \eta_i \mid \beta)]}{\partial Q}, \qquad (125)$$

$$r_i = -\frac{\partial E_{QD}[\mathrm{CVC}(pe, w, \theta_i, K(\theta_i), Q(\theta_i), \epsilon_i^D, \epsilon_i^Q, \eta_i \mid \beta)]}{\partial K}. \qquad (126)$$

The fixed fee, $T(\theta_i)$, is set so that expected profits are zero at the values of $K(\theta_i)$ and $p(\theta_i)$ that solve (125) and (126).

9.3.2. Asymmetric information model

In the asymmetric information model (Model A), the regulator recognizes that the utility may mis-report θ as higher than it really is (i.e., the utility claims to be less efficient than it really is). Consequently, the regulator constructs price, fixed fee and capital stock (as a function of θ) such that given these schedules, the utility finds it profit-maximizing to

report its true θ. The regulator picks price, the fixed fee, and the capital stock that maximize expected (with respect to the distributions of θ, ϵ^D, and ϵ^Q) consumer surplus.

To derive the Model A equilibrium, Wolak follows the approach given in Baron (1989). The first step is to determine the global truth-telling constraints. A utility with true parameter θ_x that reports θ_y earns expected profit

$$E_{QD}\big[\pi(\theta_y,\theta_x)\big] = E_{QD}\big[p(\theta_y)Q\big(p(\theta_y)\big)\epsilon^D - \mathrm{CVC}\big(\theta_x, K(\theta_y), Q(\theta_y)\big)\big] - rK(\theta_y) + T(\theta_y), \tag{127}$$

where we suppress the dependence of the minimum variable cost function (CVC) on pe, w, ϵ^Q and ϵ^D, η and β. Consequently, for any two arbitrary values θ might take for a given utility, say θ_x and θ_y, incentive compatibility requires $E_{QD}[\pi(\theta_x,\theta_x)] \geqslant E_{QD}[\pi(\theta_y,\theta_x)]$, meaning that the firm expects to earn higher profits by announcing θ_x when its true type is θ_x, than it expects to earn from announcing any other $\theta_y \neq \theta_x$. The next step is to specify the local version of this global constraint:

$$\frac{\mathrm{d}E_{QD}[\pi(\theta)]}{\mathrm{d}\theta} = -\frac{\partial E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial\theta}, \tag{128}$$

for all $\theta \in [\theta_l, \theta_h]$. Equation (128) is the local incentive compatibility condition that quantifies how rapidly the regulator must raise the expected profits of a utility as its true θ value falls (the utility becomes more efficient) in order to encourage truthful revelation. By integrating (128), one obtains the expected profit function. This implies that the expected profit function is locally decreasing in θ so that the participation constraint, which requires the firm to earn nonnegative expected profits for all values of θ, can be replaced by the single constraint that $E_{QD}[\pi(\theta_h)] \geqslant 0$, where θ lies in the interval $[\theta_l, \theta_h]$.

The regulator's optimization problem is

$$\begin{aligned}
&\max_{p(\theta),T(\theta),K(\theta)} \int_{\theta_l}^{\theta_h} \big[S_i\big(p(\theta)\big) - T(\theta)\big] f(\theta)\,\mathrm{d}\theta \\
&\text{subject to}\quad E_{QD}\big(\pi(\theta_i)\big) = E_{QD}\big[p(\theta_i)Q\big(p(\theta_i)\big)\epsilon_i^D + T(\theta_i) \\
&\qquad\qquad - \mathrm{CVC}\big(pe, w, \theta_i, K(\theta_i), Q(\theta_i), \epsilon_i^D, \epsilon_i^Q, \eta_i \mid \beta\big)\big] \\
&\qquad\qquad - r_i K(\theta_i) \\
&\qquad \frac{\mathrm{d}E_{QD}[\pi(\theta)]}{\mathrm{d}\theta} = -\frac{\partial E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial\theta}, \\
&\qquad E_{QD}\big[\pi(\theta_h)\big] \geqslant 0.
\end{aligned} \tag{129}$$

Although Wolak does not explicitly include the restrictions implied by the global truth-telling constraints, he derives restrictions on the regulatory environment and distribution of θ necessary for the price, capital, and fixed fee functions that solve (129) to also satisfy the global incentive compatibility constraints.

Note that the formulation in (129) refers specifically to the ith utility–regulator pair. Because the regulator does not know utility i's efficiency parameter, she must set p, T,

and K functions over the entire support of θ for each utility. Consequently, the regulator must solve this problem for each utility that it regulates.

Wolak (1994) presents a detailed discussion of the derivation of the first-order conditions for this optimization problem. For our purposes, we simply want to show how these first-order conditions differ from those for the Model S solution. The first-order conditions analogous to (125) and (126) for Model A are:

$$p(\theta) = \left[\frac{\partial E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial Q} + \frac{F(\theta)}{f(\theta)}\frac{\partial^2 E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial\theta\partial Q}\right]\eta_p, \tag{130}$$

$$r = -\left[\frac{\partial E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial K} + \frac{F(\theta)}{f(\theta)}\frac{\partial^2 E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial\theta\partial K}\right]\eta_K, \tag{131}$$

where η_p and η_K are the mean one multiplicative optimization errors added for same reasons given above in the discussion of the Model S solution. These two equations determine the amount of capital stock $K(\theta)$ a utility of type θ will purchase, and the price $p(\theta)$ it will be directed to charge. The demand function $Q_i(p)$ and these two equations determine the two regulatory variables $K(\theta)$ and $p(\theta)$. The fixed fee $T(\theta)$ is given by

$$T(\theta^*) = E_{QD}\big[\pi(\theta^*)\big] - E_{QD}\big[p(\theta^*)Q\big(p(\theta^*)\big)\epsilon^D + \mathrm{CVC}\big(\theta^*, K(\theta^*), Q(\theta^*)\big)\big] + rK(\theta^*) \tag{132}$$

for a utility of type θ^*. Once a utility's K is chosen and its p and T are set, its demands for L and E can be determined from the solution to the minimum operating cost problem.

These first-order conditions demonstrate that the presence of asymmetric information in the regulator–utility interaction leads to both deviations from minimum cost production and efficient output prices in the sense that price differs from marginal cost. As discussed above, this deviation from minimum cost production occurs because the firm also uses its capital stock to signal to the regulator its greater productive efficiency and therefore lower value of θ.

9.4. Econometric model

Following our procedure outlined in Section 4 for constructing structural econometric models, this section discusses the functional form of the production function $Q_i = f(K_i, L_i^*, E_i, \epsilon^Q \mid \beta)$ and derives the cost function which is used to recover an estimate of the parameter vector β. We then discuss the specification of distributions for the structural disturbances introduced into the model and derive the likelihood function. Wolak's model contains the first three types of disturbances discussed in Section 4: (1) unobserved heterogeneity in the form of the utility's private information θ_i, (2) shocks which agents in the model optimize against (ϵ^Q and ϵ^D), (3) optimization errors which allow

agents' first-order conditions to only be satisfied in expectation (η_j, $j = L, E, K, p$). Appendix A of Wolak (1994) shows that the composite errors to the structural equations are functions of these disturbances.

Wolak's choice of fairly simple functional forms for the production function and demand function allows him to impose conditions on the parameters of the underlying econometric model that guarantee a solution to the regulator's problem. More flexible functional forms for $Q_i = f(K_i, L_i^*, E_i, \epsilon^Q \mid \beta)$ would not allow this. These functional forms allow constraints on the parameters of the economic environment which guarantee the existence of a Model A solution. These functional forms allow Wolak to perform counterfactual experiments with his parameter estimates which illustrate several important empirical distinctions among Model S, Model A, and conventional estimation procedures.

Wolak uses the Cobb–Douglas production function $Q = \beta_0 K^{\beta_K} (L/\mathrm{d}(\theta))^{\beta_L} E^{\beta_E} \epsilon^Q$, where $\mathrm{d}(\theta) = \theta^{(\beta_L+\beta_E)/\beta_L}$. The demand function for the utility's output is

$$Q_d = \begin{cases} \exp(Z'b) p^{-\kappa} \epsilon^D & \text{if } p \leqslant p_{\max}, \\ 0 & \text{if } p > p_{\max}, \end{cases} \tag{133}$$

where Z is a vector of utility service area characteristics assumed shift demand, b is a parameter vector associated with Z, κ is the elasticity demand for water, and $p_{\max}$ is the price beyond which demand for the firm's output is zero.

Solving the minimum operating cost problem for this production function yields the following (conditional on K) variable cost function:

$$\begin{aligned} &\mathrm{CVC}(pe, w, K, Q, \theta, \epsilon \mid \beta) \\ &\quad = \theta \beta_0^{-\frac{1}{\beta_L+\beta_E}} K^{-\frac{\beta_K}{\beta_L+\beta_E}} \left[\left(\frac{\beta_L}{\beta_E} \right)^{\frac{\beta_E}{\beta_L+\beta_E}} + \left(\frac{\beta_L}{\beta_E} \right)^{-\frac{\beta_L}{\beta_L+\beta_E}} \right] \\ &\qquad \times w^{\frac{\beta_L}{\beta_L+\beta_E}} pe^{\frac{\beta_E}{\beta_L+\beta_E}} Q^{\frac{1}{\beta_L+\beta_E}} u, \end{aligned} \tag{134}$$

where u is a function of our previously defined disturbances ϵ^D, η_L and η_E and the parameter vector β.

Taking the partial derivative of the expected value of this cost variable function, $E_{QD}[\mathrm{CVC}]$, with respect to K and inserting it into the first-order condition for the symmetric information regulatory outcome with respect to K, yields the following unconditional variable cost (VC) function:

$$\mathrm{VC}(S) = D^* r^\alpha w^\gamma \theta^{(1-\alpha)} pe^{(1-\alpha-\gamma)} Q_d^\delta \nu. \tag{135}$$

Expressions for D^* and ν in terms of the underlying parameters of the model are given in Appendix A of Wolak (1994). The parameters α, γ and δ are defined as follows:

$$\alpha = \frac{\beta_K}{\beta_K + \beta_L + \beta_E}, \qquad \gamma = \frac{\beta_L}{\beta_K + \beta_L + \beta_E}, \tag{136}$$

$$\delta = \frac{1}{\beta_K + \beta_L + \beta_E}. \tag{137}$$

The only difference between this unconditional variable cost function and the usual Cobb–Douglas unconditional variable cost function is the presence of the utility's private information, θ.

We should emphasize that because it excludes capital costs, this is the utility's minimum variable cost function conditional on θ – not the minimum total cost function. Although it is straightforward to derive the utility's minimum total cost function from (125), Wolak departs from the tradition of estimating a total cost function for the following reason. Operating or variable costs are measured with little if any error, whereas, capital cost (the missing ingredient necessary to compute total production costs) is extremely poorly measured. Rather than complicate the analysis with potentially substantial measurement error, he instead uses the variable cost function to estimate the same parameters of the utility's production function that can be recovered by estimating a total cost function.

To derive the asymmetric information cost function, substitute the partial derivative of the expected value of the variable cost function, $E_{QD}(\mathrm{CVC})$, with respect to K into the first-order condition for the optimal capital stock given in (130). Simplifying this expression gives the following variable cost function:

$$\mathrm{VC}(A) = D^* H(\theta)^{-\alpha} \theta r^{\alpha} w^{\gamma} pe^{(1-\alpha-\gamma)} Q_d^{\delta} \nu, \tag{138}$$

where $H(\theta) = [\theta + \frac{F(\theta)}{f(\theta)}]$. The parameters α, γ and δ are as defined above.

The final step toward developing the structural econometric model is to specify distributions for all of the stochastic shocks to the econometric model. This step is needed to derive the likelihood function for the variable cost functions under the two information structures. Wolak requires that ν be lognormally distributed with $\ln(\nu) \sim N(\mu_\nu, \sigma_\nu^2)$ independent across time and utilities.

Taking the natural logarithm of both sides of (135) gives the following symmetric information logarithm-of-variable-costs equation:

$$\begin{aligned}\ln\big(\mathrm{VC}(S)\big) = {} & \xi^* + (1-\alpha)\ln(\theta) + \gamma \ln(w) + \alpha \ln(r) + (1-\alpha-\gamma)\ln(pe) \\ & + \delta \ln(Q_d) + \zeta,\end{aligned} \tag{139}$$

where $\xi^* = \ln(D^*) + \mu_\nu$ and $\zeta = \ln(\nu) - \mu_\nu$. Therefore, ζ is $N(0, \sigma_\zeta^2)$, where $\sigma_\zeta^2 = \sigma_\nu^2$. Repeating this procedure for Equation (138) yields the asymmetric information log-of-variable-costs equation:

$$\begin{aligned}\ln\big(\mathrm{VC}(A)\big) = {} & \xi^* - \alpha \ln\big(H(\theta)\big) + \gamma \ln(w) + \alpha \ln(r) + (1-\alpha-\gamma)\ln(pe) \\ & + \delta \ln(Q_d) + \zeta.\end{aligned} \tag{140}$$

The final step of the process is to define the likelihood function for each information structure. First we define notation which simplifies the presentation. Let $\Gamma^* = (\xi^*, \alpha, \gamma, \delta)'$. Define $X = (\ln(r), \ln(w), \ln(pe))'$, $q = \ln(Q_d)$, and $Y = \ln(VC)$. In this notation we can abbreviate Equations (139) and (140) as:

$$Y = \Omega_Y(X, q, \Gamma^*, \theta) + \zeta, \tag{141}$$

$$Y = \Psi_Y(X, q, \Gamma^*, \theta) + \zeta, \tag{142}$$

where $\Omega_Y(X, q, \Gamma^*, \theta)$ is the right-hand side of (139) excluding ζ and $\Psi_Y(X, q, \Gamma^*, \theta)$ is the right-hand side of (140) excluding ζ.

We now derive the likelihood function and discuss the estimation procedure for the case of Model S. Following this discussion, we describe the additional complications introduced by Model A. Under Wolak's assumptions on the functional form for the production function and the aggregate demand function the equilibrium value of q under Model S is

$$q = \big(Z', X', \ln(\theta)\big)\Lambda^* + \psi, \tag{143}$$

where Λ^* is the vector of coefficients associated with $(Z', X', \ln(\theta))$ and ψ is assumed to be joint normally distributed with ζ. Let $\rho_{\zeta,\psi}$ denote the correlation between ζ and ψ. Finally, define $\Lambda = (\Lambda^{*\prime}, \sigma_\psi^2, \rho_{\zeta,\psi})'$. Conditional on the value of θ, Equations (141) and (143) make up a triangular system of simultaneous equations. The determinant of the Jacobian of the transformation from $(\zeta, \psi)'$ to $(Y, q)'$ is one, so that the joint density of $(Y, q)'$ conditional on θ, X and Z is

$$\begin{aligned} &h_S\big(Y, q \mid \ln(\theta), \Gamma, \Lambda\big) \\ &\quad = \frac{1}{2\pi\sigma_\zeta^2\sigma_\psi^2(1-\rho_{\zeta,\psi}^2)^{1/2}} \\ &\qquad \times \exp\left[-\frac{1}{2(1-\rho_{\zeta,\psi}^2)}\big[(\psi/\sigma_\psi)^2\big] - 2\rho_{\zeta,\psi}(\psi\zeta)/(\sigma_\psi\sigma_\zeta) + (\zeta/\sigma_\zeta)^2\right], \end{aligned} \tag{144}$$

where $\Gamma = (\Gamma^*, \sigma_\zeta)'$. Note that θ enters both (141) and (143) only through $\ln(\theta)$, so that without loss of generality we can express $h_S(\cdot,\cdot)$ as a function of $\ln(\theta)$. Because θ is unobservable, to construct the likelihood function in terms of the observable variables, we must compute the density of (Y, q) given X and Z only. To obtain this density we integrate the conditional density $h_S(Y, q \mid \ln(\theta), \Gamma, \Lambda)$ with respect to the density of θ. Integrating with respect to the density of θ, yields

$$g\big(Y, q \mid X, Z, \Gamma, \lambda, F(.)\big) = \int_{\theta_l}^{\theta_h} h_S\big(Y, q \mid X, Z, \ln(\theta), \Gamma\big) f(\theta)\,\mathrm{d}(\theta). \tag{145}$$

This likelihood function is similar to Porter's regime switching model likelihood function. In Porter's case I_t is the unobserved regime indicator and in the present case θ is a continuously distributed random variable with compact support. In the same way that Porter was able to identify the density of I_t from his assumption of conditional normality of the density of equilibrium price and quantity, Wolak (1994) is able to identify the distribution of θ from the joint normality assumptions of Y and q. In addition, in the same sense that the economic structure of competitive and collusive pricing regimes was identified by the conditional normality assumption in Porter's model, the primitives of the private information regulator–utility interaction are identified by the conditional normality assumption in Wolak's model.

The construction of the likelihood function for the asymmetric information case proceeds in an analogous fashion, with the major complication being the presence of $H(\theta)$, which is a function of both $f(\theta)$ and $F(\theta)$ in both regression equations. The conditional density of $(Y, q)'$ given θ, X and Z under Model A takes the same form as for Model S with Equation (141) replaced by Equation (142) and the log-output Equation (143) replaced by the following equation:

$$q = \big(X', Z', \ln(\theta), \ln\big(H(\theta)\big)\big)\Phi + \psi, \tag{146}$$

where Φ is the vector of coefficients associated with $(X', Z', \ln(\theta), \ln(H(\theta))')$.

The conditional distribution of Y and q given Z, X, and θ for this information structure, $h_A(Y, q \mid X, Z, \theta)$, depends on θ through both $\ln(\theta)$ and $\ln(H(\theta))$. To construct the likelihood in terms of only observables, we integrate this conditional density with respect to $f(\theta)$ over the interval $[\theta_l, \theta_h]$.

For both Model S and Model A, conventional maximum likelihood estimation procedures can be applied to compute the coefficient estimates and their standard errors.

9.5. Estimation results

A major goal of the empirical analysis is to recover characteristics of production process, and in particular, the returns to scale in production, accounting for the impact of the utility's private information. Wolak finds that applying conventional minimum cost function Cobb–Douglas estimation techniques, the returns to scale estimates obtained that are implausibly high, with cost elasticities with respect to output estimates as high as 0.77, which means that a 10 percent increase in output only increases total costs by 7.7%. Other estimates were even lower. However, applying the maximum likelihood estimation techniques outlined above for the Model S and Model A solutions, Wolak finds cost elasticities with respect to output greater than 1, for both the Model S and Model A estimates, which implies slight decreasing returns to scale in production, although the null hypothesis of constant returns to scale cannot be rejected. This dramatic difference in returns to scale estimates points out the importance of controlling for this unobserved firm-level heterogeneity in productive efficiency when attempting to recover consistent estimates of the characteristics of the regulated firm's production process.

Wolak is also able to recover estimates of $F(\theta)$, which determines the form of the optimal regulatory contract under asymmetric information. Armed with this information he is able to compute the following counterfactuals for each point in his dataset using the Model A parameter estimates. First, he computes the ratio of total operating costs under a Model A solution versus the Model S solution holding constant the level of output produced by the firm under both scenarios. This answers the question of how much less costly, in terms of variable costs, it is to produce a given level of output under the Model A versus Model S versions of the regulatory process. Wolak also performs this same counterfactual for total production costs and finds that in terms of total production costs, the same level of output costs approximately 5–10 percent more to provide

under the Model A regulatory process relative to the Model S regulatory process. These distortions from minimum cost production occur because the more efficient firms find it profitable to signal their superior productive efficiency to the regulator through their capital stock choice.

Wolak also computes the welfare cost to consumers from asymmetric information by comparing the market-clearing level of output under the Model A solution versus the Model S solution for the same values of the input prices and θ. He finds the output level produced under the Model A solution is roughly 20 percent less than the level of output under the Model S solution for the Model A parameter estimates, which indicates a significant welfare loss to consumers associated with asymmetric information.

In an attempt to see whether Model A or Model S provides a statistically superior description of the observed data, Wolak performs a nonnested hypothesis test of Model A versus Model S. He finds that Model A provides a statistically significantly superior description of the observed data relative to Model S. As discussed in Section 3, this does not validate Model A as the true model for the regulatory process. It only states that for the same functional forms and economic primitives, the strategic interaction implied by Model A provides a statistically superior description of the observed data.

9.6. Further extensions

There are variety of directions for future research in this area given the enormous number of competing theoretical models of the private information regulator–utility interaction. Sorting through the empirical implications of these models across a variety of regulated industries would help to focus future theoretical and empirical research in this area. Recent work in this area includes: Dalen and Gomez-Lobo (1997) who study the impact of these incentive contracts in the Norwegian Bus Transport Industry and Gagnepain and Ivaldi (2002) who assess the impact of incentive regulatory policies for public transit systems in France.

10. Market structure and firm turnover

So far we have discussed IO models in which the number of market participants (e.g., firms or bidders) is given. IO economists have recently devoted considerable energy toward modeling how changes in market structure can affect the extent of competition in a market. In particular, the theoretical literature has explored two related questions:

1. "How many competitors are needed to insure effective competition?" and
2. "What factors encourage firms to enter markets?"

Theoretical answers to these questions often hinge delicately on the assumptions made about firms' costs, market demand and firms' conjectures about competitors' behavior. Unfortunately, there are very few structural econometric models that would allow one to identify the empirical relevance of demand, cost and strategic explanations. In large part

this is because competition models in which the number of participants is endogenous are complicated and difficult to solve.

Only recently have empirical researchers begun to make progress in developing structural econometric models that can speak to specific strategic models of entry and entry deterrence. In this section we outline some of the econometric issues associated with modeling the number of firms in oligopolistic markets. Again, our intent is not so much to survey the literature as to show what one can learn from information about the number and identities of firms in a market.[32] We shall see that while structural models of entry, exit and market structure raise many of the modeling issues discussed in Sections 5–9, there are also new issues.

10.1. Overview of the issues

Sections 5–7 showed how economists have used information about the joint density of prices and quantities $f(P, Q \mid X, Z) = f(P_1, \ldots, P_N, Q_1, \ldots, Q_N \mid X, Z)$ to recover information about firms' demand curves and costs. In general, the conditional density $f(\cdot)$ is a statistical object, and a high-dimensional one at that. In practice this means that it would be hopeless to try and estimate a $2 \times N$ conditional joint density nonparametrically from market-level data. While going to consumer-level data can improve inferences, in general it will be extremely difficult to obtain the representative consumer-level datasets necessary to estimate flexible and yet precise estimates of firm-level demands. These observations suggest that considerable economic structure will have to be introduced if one is to obtain meaningful estimates of firms' demands and costs.

The literatures discussed in Sections 5–9 presume that the number of firms is exogenous. One consequence of this assumption is that N enters objects such as $f(P, Q \mid X, Z)$ as a conditioning variable rather than something to be explained. One way to make N endogenous is to imagine that each market has the same $M > N$ potential entrants. Each of these potential entrants makes a discrete decision whether or not to enter. The conditional density of the market data and these entry decisions is $f(P_1, \ldots, P_M, Q_1, \ldots, Q_M, a_1, \ldots, a_M \mid X, Z, W, M)$. Here, the a_i are zero-one indicators for whether or not potential entrant i has entered and W are any new conditioning variables.

This expression makes it easy to appreciate why many studies do not make N endogenous. First, there are many different collections of the a_i that yield the same N. In principle, the researcher might wish to explain not just N but why a particular ordering of the a_i was obtained. Second, because the dimensionality of $f(\cdot)$ has gone up considerably, it becomes even more difficult to estimate nonparametrically. For example, it seems unlikely that a researcher would have a large sample of markets that have the same number of potential entrants M. Finally, the form of $f(\cdot)$ may differ with the identities of each entrant.

[32] For a more complete discussion see Berry and Reiss (in press).

Because nonparametric methods are impractical, the researcher will have to impose economic structure to get anywhere. In particular, now the researcher will have to add equations that explain each of the a_i. These conditions must explain why some but not other potential entrants entered the market.

Our discussion so far has dealt with more obvious complications introduced by making N and the identities of entrants endogenous. There are less obvious complications as well. Two of the most critical are that: (1) the underlying theory may deliver ambiguous predictions about which firms will enter in equilibrium; and, (2) the underlying theory may deliver no (pure-strategy) predictions about which firms will enter in equilibrium. These are new complexities, ones we did not really see in Sections 5–9. Before we explore their significance for structural modeling, it is useful to back up and provide a broader sense of the types of economic issues that one might hope to address with structural models of market concentration and competition.

10.1.1. Airline competition and entry

Since the deregulation of US passenger airline markets in the late 1970s, travelers and economists have speculated about whether sufficient competition exists in different city-pair markets.[33] One does not have to look far to understand why. Travelers routinely encounter wide disparities in an airline's fares (per seat mile) over time, across routes and even for seats on the same flight. Despite this considerable variation in a given airline's fares, there appears to be much less variation in fares across competing carriers. Industry critics contend that such patterns are obvious evidence of ineffective competition. They also argue that high concentration on some individual city-pair routes contributes to the problem. Some industry advocates argue the opposite. They contend that fare matching is evidence of competition, and that fare differences at worst reflect price discrimination. Some also claim that high concentration is evidence of economies of scale and route density, and that entry (or the threat of entry) of small upstart carriers is enough to insure effective competition.

These two views provide a challenge to IO economists, and there have been many attempts to distinguish between them. To delve deeper, it is useful to imagine that we have data (consistent with the US experience) indicating that short haul routes between small cities tend to be highly concentrated and have high (per seat mile) fares. The technological and demand explanation for this correlation is that the costs of service on these routes is high relative to demand. Thus, some routes will have so little demand relative to costs, that at most one firm can profitably serve the market. This one firm would behave as a monopolist and charge high prices to recover its costs. The anti-competitive explanation for the observed correlation is that high concentration and fares are the result of strategic behavior. For example, even if the small market could support

[33] See for example Borenstein (1992), Brueckner, Dryer and Spiller (1992), Morrison and Winston (1996), Ott (1990), and Windle (1993).

many carriers, dominant carriers can convince potential entrants that entry would be met with stiff competition.

Can we distinguish between these explanations? Our answer is: given the current state of the theory, econometric models and data, we cannot generally. The main reason is that much of what the theory points us toward is unobservable. Researchers do not observe the marginal and fixed costs that are central to technological explanations. We also do not observe potential entrants' expectations about incumbent behavior, which are central to strategic explanations. Does this mean we cannot learn anything from a structural model of market structure? The answer to this is no.

What we can imagine doing in principle is building structural models that would examine how alternative competitive models fit the data. For instance, we might begin in the spirit of the models in Sections 5–7 by writing down functional forms for city-pair demand, and firms' fixed and variable costs. This is not, however, as easy as it sounds. Prior studies have documented that airlines' costs of service depend in complex ways not only on route-specific factors, such as miles traveled, airport fees, etc., but also on network and fleet characteristics (e.g., whether the plane will carry passengers beyond a city or transfer passengers at a hub and code-sharing agreements). Nevertheless, we might attempt a parametric model of demand and costs. At that point, unlike most of the models in Sections 5–7, we would have to grapple with the problem that the number of carriers in a market is endogenous: it is affected by demand and supply conditions. We therefore also have to model how fixed and marginal costs impact the number of firms in the market (and possibly the identities of those firms).

Here, we encounter tricky specification issues. Economic theory suggests that to model the number of firms we need to model why (and possibly which) firms did not enter. But this involves modeling potential entrants' expectations about what would happen after entry, something we never observe. Moreover, because the same carriers compete with each other in other markets, we may have to model how actions in any one market affect outcomes in other markets.

At this point, it might seem that a complete structural model of airline competition is hopeless. There is, however, something that we can learn with the right data. The critical events that tell us something about competition and market structure are instances of entry and exit. Consider, for example, our sample of small markets. In principle, we observe some city-pair markets in which there is no (direct) service, others in which there is a monopoly, a duopoly, and so on. If (and this is an important if) we can control for factors that might lead to cost of service and demand differences across markets, then we can ask how much demand does it take to support at least one carrier. This level of demand tells us something about a single carrier's fixed and marginal costs relative to demand. We can then compare this level of demand to what it takes to support a second firm in the market. This level of demand tells us more about costs and potentially behavior. Suppose, for instance, we do not observe a second carrier enter a city-pair market until demand is roughly twenty times what it takes to support a single carrier. One's intuition is that if the second carrier has the same costs and product as the first,

that this difference must reflect pessimism on the part of the second carrier as to value of entering a monopoly market.

It is this type of intuition that structural models of the number of firms, or entry and exit seek to make more precise. That is, the goal of a structural model is to show how changes in population and other exogenous market conditions affect the (apparent) ability of potential entrants to cover costs. The primary value of a formal model is that it makes clear what economic and stochastic assumptions are necessary, given the available data, to isolate differences between firms' costs and the expectations they have about post-entry competition.

10.2. An economic model and data

Our airline example makes three points that are worth emphasizing. First, debates about the competitiveness of markets often hinge on assumptions about what determines a market's structure (e.g., the number of firms). Second, some of the most critical factors affecting the ease of entry and exit are unobservable (e.g., firms' fixed and marginal costs, and expectations about post-entry competition). Third, while we can potentially use structural models to draw inferences about the unobservables present in IO theories, these models, like all structural models, will contain untestable assumptions. These assumptions may be too numerous to be credible.

An important corollary to this third point is that the form of the data available will have an important impact on what we can estimate. In our airline example, for instance, we might have data on a cross section of similar city-pair markets or time series data on the same market over time. Both of these data sets raise modeling issues. In cross-section data we have to worry about changes in the identity and number of potential entrants across markets. We may also have to worry that the behavior of firms in one market may affect their behavior in other markets. While time-series data have the advantage of holding constant market-specific conditions, researchers must again worry that the firms' decisions may be linked through time. When they are, it makes sense to model firms' decisions using dynamic games. While some progress has been made in formulating and solving such games, to date their computational demands have largely made them impractical for empirical work. As a consequence, almost all structural market structure models are static.

Most empirical work in this area has tended to rely on cross-section data. As such they focus on modeling which firms are producing, as opposed to firm turnover; i.e., which firms are entering or exiting. In a typical cross-section application, a researcher might have data on

1. the number of potential entrants into each market, M;
2. the entry decisions of each potential entrant: $a = (a_1, a_2, \ldots, a_N)$;
3. market-specific information X (e.g., market size); and
4. firm-specific information, $Z = (z_1, z_2, \ldots, z_M)$ (e.g., identities and product characteristics).

In addition, in an ideal application the researcher may also observe the prices and quantities of actual entrants: $P_1, \ldots, P_N$ and $Q_1, \ldots, Q_N$.

In an ideal setting, the structural modeler would like to use this information to estimate firm-level demand and cost specifications, such as those discussed in Sections 5–8. Unlike these previous models, however, assumptions about firms' fixed costs will now play an important role in these models, as fixed costs help determine which set of firms will produce. Additionally, assumptions about the timing of firms' decisions and the amount of information they possess become critical. These assumptions are important because, unlike in previous models, they have a critical impact on whether the empirical model has a pure-strategy equilibrium and whether any pure-strategy equilibrium is unique. In what follows, we use a series of models advanced by Bresnahan and Reiss (1991a, 1991b) to highlight some of these issues and the strengths and weaknesses of structural models.[34]

Bresnahan and Reiss develop econometric models to explain the number of sellers in several different localized product markets (such as dental services, new car dealers and movie theaters). For each product, they model how the number of sellers in a town varies with the town's population, and other demand and cost variables. The goal of their work is to understand how technological, demand and strategic factors affect market structure and competition. Like the airline example, they propose to do this by estimating how much demand it takes to support different numbers of firms. Unlike the airline example, however, the authors only have information on the number of firms in each market and their identities $a = (a_1, \ldots, a_M)$; they do not have price or quantity information. Thus, absent a structural model, the best they can do is summarize the conditional joint distribution of entry decisions given industry and firm characteristics. Such an approach is not that dissimilar from that taken in Dunne, Roberts and Samuelson (1988). When developing a structural model, Bresnahan and Reiss must take into account the fact that entry and exit are discrete events. Thus, their structural models will not typically involve marginal conditions, such as those used in the models of Sections 5, 6 and 7. Instead, they must rely on threshold conditions for entrants' unobserved profits.

The threshold conditions that Bresnahan and Reiss use come from simple static, perfect-information entry games. An example of such a game is the standard two-firm, simultaneous-move entry game. The payoffs to the players in this game are:

	Stay out ($a_2 = 0$)		Enter ($a_2 = 1$)	
Stay out ($a_1 = 0$)	$\Pi_1(0,0)$	$\Pi_2(0,0)$	$\Pi_1(0,1)$	$\Pi_2(0,1)$
Enter ($a_1 = 1$)	$\Pi_1(1,0)$	$\Pi_2(1,0)$	$\Pi_1(1,1)$	$\Pi_2(1,1)$

where the $\Pi_k(a_1, a_2)$ represent the profits firm k earns when firm 1 plays a_1 and firm 2 plays a_2 (a zero denotes the action "Stay Out" and a one denotes "Enter"). In most

[34] See also the work of Berry (1992) and other references cited in Berry and Reiss (in press).

textbook examples, the numbers in the payoff matrix are hypothetical. The economist then adds assumptions about players' information and a solution concept.

Bresnahan and Reiss' structural models build on this strategic representation of an entry game. Their econometric models postulate that the researcher observes the players' equilibrium action(s) in each sample market (e.g., $a_1 = 0$ and $a_2 = 1$) but does not observe the firms' economic profits (the $\Pi_k(0, 1)$). The logic of their models is to use a specific equilibrium solution concept to work backward from the observed equilibrium action(s) to statements about unobserved profits. Thus, the "structure" in their structural model are the economic and stochastic assumptions that allow them to go from discrete data to statements about continuous-valued profits. It should not be too surprising given our discussions in Sections 5–9, that Bresnahan and Reiss will have to introduce considerable structure in order to draw inferences about firm profits and behavior from discrete outcomes.

10.3. Modeling profits and competition

To understand the process by which Bresnahan and Reiss work from firms' observed actions back to statements about firms' unobserved profits, and to see what one can hope to estimate, it is useful to work with a specific entry model. To keep matters simple, imagine that we are modeling the number of symmetric firms, N, that produce a homogeneous good. The goal of the empirical analysis is to use the information in the zero-one entry indicators $a_1, a_2, \ldots, a_M$ of the $M \geqslant N$ potential entrants to draw inferences about firms' profit functions, i.e.,

$$\Pi_k(a_1, a_2, \ldots, a_M, X, Z, W, \theta). \tag{147}$$

Here X, Z, and W represents exogenous observables affecting demand and costs, and θ represents parameters of the profit function (e.g., demand and cost function parameters) that we wish to estimate. While the firms' profit functions could in principle include prices and quantities, Bresnahan and Reiss do not have this information. They thus are forced to work with profit functions where these endogenous variables have been substituted out.

The first step in the modeling process is to use assumptions about demand, costs and how firms compete to derive the functional form of Equation (147). Here Bresnahan and Reiss are helped by the presumption that if a potential entrant does not enter, it likely will earn zero profit – regardless of what the other potential entrants do. If firm i does enter, its profits depend on the number of other firms that enter (as summarized in the a_j). The exact way in which the number of other firms affects profits depends on what one assumes about demand, costs and competition. If, for example, firms have the same constant marginal cost c, have fixed costs of F, compete as Cournot competitors, and market demand is $p = \alpha - bQ$, then one can show

$$\Pi_k(a_1, a_2, \ldots, a_M, Z, \theta) = b\left(\frac{S}{\sum_{j=1}^{M} a_j + 1}\right)^2 - F, \tag{148}$$

where $S = (\alpha - c)/b$ is a measure of the potential size of the market. For firm i to have entered along with $N - 1$ other firms it must be the case that $\Pi_i \geqslant 0$. Similarly, if there is free entry, then it must be that the $(N + 1)$st entrant found it unprofitable to enter. These two bounds imply

$$\frac{S^2}{(N+1)^2} \geqslant \frac{F}{b} \geqslant \frac{S^2}{(N+2)^2}.$$

These inequalities provide useful information. For instance, if we know or could estimate the size of the market S and the slope of demand b, then we can place a bound on firms' unobserved fixed costs. While it is plausible to imagine having external measures of the market's size, S, it is much less likely one would have prior information about b. One solution would be to use price and quantity data to estimate b, yet this is exactly the problem that Bresnahan and Reiss have – they do not have price and quantity information.

The question then is what can one infer about demand and cost conditions from a cross section of markets? Bresnahan and Reiss' idea is to use information on the number of firms in very small to very large markets to estimate a sequence of so-called entry thresholds. These thresholds are a simple transformation of the market sizes $S_1, S_2, \ldots$ above, where S_i represents the size of the market just needed to support i firms. While the entry threshold levels are of limited use, their ratios are revealing. For example, if we take the ratio of the duopoly to the monopoly entry threshold assuming firms are Cournot competitors we get

$$\frac{S_2^2}{S_1^2} = \frac{9}{4} = 2.25. \tag{149}$$

That is, we should observe a second firm entering at 2.25 the size of the market required to support one firm. Similar calculations can be done for entry threshold ratios involving higher numbers of identical firms.

Of course, we need not observe the estimated (or observed) duopoly-monopoly threshold ratio equal to 2.25 (or the higher-order ratios consistent with this symmetric Cournot model). The question then is what should we infer? The answer is that economic theory can provide some suggestions. We can consider, for example, what happens when we change the assumption about how the duopolists compete. If the second entrant expects the monopolist to collude with it after entry, then the duopoly to monopoly ratio would equal 2.0. The three-firm to monopoly entry threshold ratio would be 3.0, and so on. Alternatively, if the second firm expected perfect competition (or Bertrand competition) post entry, we would never observe the second firm enter this natural monopoly. Thus, we can see that the degree of competition affects the entry threshold ratio. While we might be tempted to think the entry threshold ratio then is indicative of the degree of competition, with larger ratios suggesting more competition post entry, this is only true if we maintain our other assumptions. If, for example, we had used a quadratic cost function with increasing marginal costs, we also would see

changes in the entry threshold ratios as minimum efficient scale changes [see Bresnahan and Reiss (1991a)].

This last point brings us back to a point we made in the introduction: inferences in structural models typically depend heavily on maintained functional form assumptions. We often do not have the data to test these assumptions. In this application, for example, the absence of price and quantity data considerably limit what we can infer. Does this suggest that this structural model has little value because we have to make untestable assumptions? Our answer is no. The model has value because it makes clear what one can and cannot infer from the data. It also points future research toward what it is that one needs to observe to draw sharper inferences.

10.4. The econometric model

Our discussion so far has largely been based on an economic model with symmetric firms. We have yet to introduce stochastic assumptions or discuss the more realistic cases where there are observed and unobserved differences among firms. These additions introduce further complexities.

Recall that the data Bresnahan and Reiss have are the number of potential entrants M, the number (and possibly the identities) of the actual entrants, and demand and cost variables. Starting from primitive demand and cost function assumptions, they build a model of firms' equilibrium profits, which consists of a variable profit and a fixed cost term

$$\bar{\Pi}_k(a, Z, \theta) = \mathrm{VP}_i(a, Z, \theta) - F_i(a, Z, \theta). \tag{150}$$

Here, a is a vector describing the M potential entrants' entry actions, VP denotes variable profits, F fixed costs and i subscripts potential entrants. Although this expression depends on observable variables, the econometrician does not typically observe everything the firm does. Following the discrete choice literature popularized by McFadden, Heckman, and others, we might simply add an error term, ϵ, to profits to account for what we do not observe. Notice, however, that by assuming that the error is additive, we have placed structure on what it is about profits that the econometrician does not observe. Specifically, whatever it is that the econometrician does not observe, it enters the firms' optimal choices of prices and quantities in such a way that we obtain an additive error in Equation (150). What types of unobservables do and do not fit this specification? If we assume that the firms have unobserved differences in their constant marginal costs, then we will not obtain an additive error specification. On the other hand, if we assume that firms have different fixed costs, then we will. (This is because the marginal conditions for prices or quantities do not depend on the unobservable fixed cost.) Thus, while it is possible to justify the unrestricted additive structure in (150), it may make more economic sense to entertain alternative stochastic specifications for profits.

Assuming that the unobserved portion of profits is additive, we are now in a position to write down expressions for the equilibrium threshold conditions on firm profits. Following the discrete choice literature, we might consider modeling entry as the event that

the firm i's latent profits exceeds 0, or

$$\mathrm{VP}_i(a, Z, \theta) - \tilde{F}_i(a, Z, \theta) \geqslant \epsilon_i(a), \tag{151}$$

where the tilde above fixed costs denotes fixed cost up to an additive mean zero error. This model looks like a standard threshold condition in a conventional discrete choice model. The key difference is that the threshold conditions in the entry model contain the endogenous a_i variables. In other words, unlike in the standard discrete choice model, here agents' discrete decisions are interrelated. We therefore have to model simultaneously the N potential entrants' threshold conditions. This is the source of additional complications.

There is some precedent in the discrete choice literature for threshold conditions that include dummy endogenous variables (the a_i). For example, the household labor supply literature sometimes descriptively models the dependence of a household head's labor supply decision on their spouse's labor supply decision. Amemiya (1974) and others have studied the econometric properties of latent variable models that include dummy endogenous variables. Heckman (1978) introduced a systematic formulation of linear dummy endogenous variable models and discussed a variety of econometric issues associated with the formulation and estimation of such models. In particular, he and others have noted that arbitrary specifications of dummy endogenous variable models can lead to "coherency" and identification problems.

Bresnahan and Reiss showed that one could use the economic structure of discrete games to produce structural choice models with Heckman's econometric structure. Moreover, the identification issues that arise in Heckman's models often have natural economic interpretations. To see some of the connections, let us return to the normal form entry game above. Recall that the idea of Bresnahan and Reiss is to draw inferences about the unobserved payoffs from the observed equilibrium actions of the entrants. To link the observed actions to the payoffs, we employ an equilibrium solution concept. An obvious one to employ in analyzing an entry game is that of a Nash equilibrium. An outcome $\{a_1^*, a_2^*\}$ of the entry game is a Nash equilibrium if

$$\Pi_1\big(a_1^*, a_2^*\big) \geqslant \Pi_1\big(a_1, a_2^*\big) \quad \text{and} \quad \Pi_2\big(a_1^*, a_2^*\big) \geqslant \Pi_2\big(a_1^*, a_2\big) \tag{152}$$

for any a_1 and a_2. To make clear the connection between the Nash equilibrium outcomes and payoffs, we can rewrite the two-by-two entry game as:

	Stay out ($a_2 = 0$)	Enter ($a_2 = 1$)
Stay out ($a_1 = 0$)	$\Pi_1(0,0) \quad \Pi_2(0,0)$	$\Pi_1(0,1) \quad \Pi_2(0,0) + \Delta_0^2$
Enter ($a_1 = 1$)	$\Pi_1(0,0) + \Delta_0^1 \quad \Pi_2(1,0)$	$\Pi_1(0,1) + \Delta_0^1 + \Delta_1^1 \quad \Pi_2(1,0) + \Delta_0^2 + \Delta_1^2$

where the Δ's represent the incremental profits to each firm of entry. From the definition of a Nash equilibrium and the above payoff matrix we can deduce

$$a_1 = 0 \quad \Longleftrightarrow \quad \Delta_0^1 + a_2 \Delta_1^1 \leqslant 0,$$

$$a_2 = 0 \quad \Longleftrightarrow \quad \Delta_0^2 + a_1 \Delta_1^2 \leqslant 0. \tag{153}$$

These conditions link the observed actions to profits. Specifically, they tell us that all that the econometrician can infer from the observed equilibrium actions are statements about the Δ terms. In the case of a Nash equilibrium, we see this means that the econometrician cannot estimate $\Pi_1(0, 1)$ and $\Pi_2(1, 0)$, which are the profits the firms earn when it is out of the market. This makes perfect sense, as we can only learn about profits when a firm enters. To understand what we can estimate, it is useful to analyze the Δ's. The Δ_0^i term are the incremental profits that firm i earns in a monopoly. We might naturally think of this incremental profit as monopoly variable profits minus fixed costs, net of opportunity costs. The Δ_1^i terms are the profits that firm i gains (loses) relative to its incremental monopoly profit when it enters its competitor's monopoly market. This profit is most naturally thought of as the loss in variable profit from moving from a monopoly to a duopoly.

From assumptions about the structure of demand and costs, we can relate the incremental profit terms to underlying demand and cost variables and parameters. For example, in the symmetric linear demand and cost Cournot example, where $\Pi_i(0, 0) = 0$ we have

$$\begin{aligned} \Delta_0^i &= \frac{(\alpha - c)^2}{4b} - F = g(\alpha, c) - F, \\ \Delta_1^i &= \frac{5(\alpha - c)^2}{36b} = h(\alpha, c). \end{aligned} \tag{154}$$

Knowing this relationship between the Δ's and the underlying economic parameters, we can proceed to add error terms to the model to generate stochastic specifications. Assuming $F_i = F + \epsilon_i$ gives the following latent variable system

$$a_i = \begin{cases} 1 & \text{if } y_i^* = g(\alpha, c) - F + a_j h(\alpha, c) - \epsilon_i \geqslant 0, \\ 0 & \text{if } y_i^* = g(\alpha, c) - F + a_j h(\alpha, c) - \epsilon_i < 0, \end{cases} \tag{155}$$

for $i = 1, 2$ and $i \neq j$. This system bears a resemblance to Heckman's (1978) linear dummy endogenous variable systems. For instance, if we ignore the demand and cost parameters in $g(\cdot)$ and $h(\cdot)$, assume Δ_1^i is a constant, and $\Delta_0^i = X\beta_i$, where X is a vector of observable variables and β_i is a vector of parameters, then we obtain the linear dummy endogenous variable system

$$a_i = \begin{cases} 1 & \text{if } y_i^* = X\beta_i + a_j \delta - \epsilon_i \geqslant 0, \\ 0 & \text{if } y_i^* = X\beta_i + a_j \delta - \epsilon_i < 0. \end{cases} \tag{156}$$

Amemiya, Heckman, Maddala and others have noted we cannot estimate the above systems in general if the errors have unbounded support. The reason for this is that the reduced form is not always well defined for all values of the errors. Bresnahan and Reiss show that this econometric problem has a natural economic interpretation: namely, it is indicative of two types of problems with the underlying game. First, if the errors are unrestricted. the underlying game may have multiple pure-strategy equilibria.

Second, the underlying game may have no pure-strategy equilibria. These existence and uniqueness problems cause havoc with pure-strategy reduced forms.

One proposed solution to these problems is to assume that the model is recursive. This econometric solution, however, has unattractive economic implications for an entry game. Specifically, it amounts to assuming that a competitor's entry into a monopoly market does not affect the monopolist's profits. Thus, while this assumption is computationally attractive, it is economically and empirically unrealistic.

Bresnahan and Reiss go on to suggest how one can impose restrictions on profits that remove existence problems. They also suggest a solution for the nonuniqueness problem, which is to aggregate the nonunique outcomes (in this case the nonunique outcomes occur when one firm or the other firm could be a profitable monopolist) to obtain an economic model *of the number of firms in the market*, rather than a model *of which firms are in the market*. Bresnahan and Reiss also explore how changing the solution concept for the entry model changes the econometric structure of the game. The main one they explore is how changing the game from simultaneous-move Nash to sequential-move Stackleberg. In the latter case, the entry game generically has a unique equilibrium. The econometric model of this equilibrium also has a threshold interpretation, but it is more complicated than the simple linear structure above.

10.5. Estimation

Turning now to estimation, Bresnahan and Reiss (1991a) propose maximum likelihood methods for estimating the parameters of profits. In their empirical work, they focus on estimating models where the number of potential entrants is small. A key assumption in their work is that they actually know the number of potential entrants, and therefore the number of threshold conditions to impose. In much of their work, they ignore systematic differences in firms' profits and focus instead on modeling the number of firms that will enter geographically distinct markets. In particular, Bresnahan and Reiss assume that the demand for the products they look at is proportional to a town's current and future population size, and that the per capita demands for these products does not depend on population. This allows them to express market demand as $Q = D(Z, P)S$, where S is the "size" of the market. To simplify the analysis, Bresnahan and Reiss assume that sellers are the same, apart from potential differences in fixed costs.

Using these assumptions, Bresnahan and Reiss derive expressions for equilibrium monopoly and duopoly profits as a function of the size of the market S, other demand variables and cost variables. A key observation is that the size of the market S enters linearly into firm profits. Assuming there are only two possible entrants, firm 1 has post-entry profits

$$\Pi_i(1, a_2) = \big(g(Z, \beta) + a_2 h(Z, \delta)\big)S - F(a_2) - \epsilon. \tag{157}$$

From this relation, Bresnahan and Reiss identify entry thresholds for a monopolist and a duopoly. That is, the entry thresholds equal

$$S(a_2) = \frac{F(a_2) - \epsilon}{g(Z, \beta) + a_2 h(Z, \delta)}. \tag{158}$$

The entry thresholds are of interest because they tell us something about unobserved fixed costs relative to the variable profit parameters. While in principle, Bresnahan and Reiss should motivate the functions $h(Z, \delta)$ and $g(Z, \beta)$ from a specific model of demand and variable costs, in their empirical work they assume that these functions are linear in the Z variables (or constants). Bresnahan and Reiss make these assumptions both to simplify estimation and because they cannot easily separate cost and demand variables.

In most of their work, Bresnahan and Reiss focus on estimating ratios of entry thresholds. In their model, the ratio of the monopoly to the duopoly entry threshold equals:

$$\frac{S(1)}{S(0)} = \frac{F(1)}{F(0)} \frac{g(Z, \beta)}{g(Z, \beta) + h(Z, \delta)}. \tag{159}$$

This expression shows that the ratio depends on the extent to which the second entrant has higher fixed costs than if it were a monopolist and the extent to which duopoly profits are less than monopoly profits (here $h(Z, \delta) < 0$). Bresnahan and Reiss estimate the left-hand side by first estimating the parameters of the profit functions (150) and then forming the ratio (159). They then draw inferences about competition based on maintained demand and cost assumptions, much as we have discussed above. For example, they observe that entry threshold ratios in several different product markets are not dramatically different from that implied by a model where firms act as Cournot competitors. Again, however, their inferences about product market competition rest heavily on their assumptions about demand and costs, and they only explore a limited set of alternative demand and cost assumptions.

10.6. Epilogue

In Section 4 we stated that a structural modeling exercise should not go forward without a clear justification, in terms of economically meaningful magnitudes that can be estimated, for the many untestable assumptions necessary to specify and estimate a structural model. In this case, the justification for the structural model is its ability to recover estimates of the entry thresholds and the fixed costs of entry from the number of firms in a market. Neither of these magnitudes are directly observable and thus can be inferred after the researcher has made assumptions about the form of demand and firm-level costs, including entry costs. In contrast to the literature described in Sections 5 through 7 that uses market prices and quantities, with fewer observable market outcomes, these models rely more heavily on functional form and distributional assumptions to recover magnitudes of economic interest.

A number of researchers have extended Bresnahan and Reiss' models and explored alternatives [see Berry and Reiss (in press)]. In many respects these models share a common feature: to draw economic inferences from qualitative data on entry and exit, they have to impose considerable economic structure and in many cases sacrifice realism to obtain empirically tractable specifications. So what does this say about IO economists' progress in developing structural models of oligopolistic market structure? The bad news is that the underlying economics can make the empirical models extremely complex. The good news is that the attempts so far have begun to define the issues that need to be addressed. They also have clarified why simple probit models and the like are inadequate for modeling entry and exit decisions.

11. Ending remarks

More than fifty years ago, members of the Cowles Commission began a push to estimate empirical models that combined economic models with probability models. They labeled this enterprise econometrics. In the intervening years, some economists have come to think of econometrics as high-tech statistics applied to economic data. That is, that econometrics is a field that mainly focuses on the development of statistical techniques. While this may be true of some of econometrics, much of the Cowles Commission's original vision is alive and well. In this chapter, we have tried to provide a sense of how structural modeling proceeds in industrial organization. We used "structural econometric modeling" as opposed to "econometric modeling" in our title to emphasize that an application's setting and economics should motivate specific probability models and estimation strategies, and not the other way around.

We began by comparing nonstructural or descriptive, and structural models. We should emphasize once more that we see great value in both descriptive and structural models. IO economists, for example, have learned much about the sources of competition from case studies of competition in specific industries. Our introductory sections tried to provide a sense of the benefits and costs associated with developing and estimating descriptive and structural models. An important benefit of a structural model is that it allows the researcher to make clear how economics affects the conditional distribution of the data. For example, we can always regress market quantity on price, but this does not necessarily mean we have estimated the parameters of a market demand function. To know whether we have or have not, we need to be clear about supply and the sources of error in the estimating equation.

While economic theory can help guide the specification and estimation of economic quantities, there is no simple recipe for developing structural econometric models. There are a variety of factors that make structural modeling difficult. First, economic theories often are sufficiently complex that it is difficult to translate them into estimable relations. In this case, structural modelers who opt to estimate simpler models often are subject to the criticism that their models are too naive to inform the theory. Second, structural modelers often lack data on all of the constructs or quantities in an economic theory.

The absence of relevant data can considerably complicate estimation and limit what it is that the researcher can estimate with the available data. Third, economic theory rarely delivers all that the structural modeler needs to estimate a model. Much is left to the modeler's discretion. The structural modeler typically must pick: functional forms; decide how to measure theoretical constructs; decide whether to include and how to include variables not explicitly part of the theory; how to introduce errors into the model; and decide on the properties of errors. Each of these decisions involve judgments that cannot be tested. Thus, these maintained assumptions need to be kept in mind when interpreting structural model estimates, parameter tests and performing counterfactual calculations.

In our selective tour, we have tried to provide a sense of how IO researchers have dealt with some of these issues. Our intent was not to be a comprehensive review of all that has been done on a particular topic, but rather to provide a vision for some of the general modeling issues IO researchers face in linking IO theories to data. We hope that our chapter has conveyed a sense of progress, and also a sense that much remains for IO economists to explore.

References

Ackerberg, D., Rysman, M. (2005). "Unobserved product differentiation in discrete choice models: Estimating price elasticities and welfare effects". RAND Journal of Economics 36 (4), 771–788.

Amemiya, T. (1974). "Multivariate regression and simultaneous equation models when the dependent variables are truncated normal". Econometrica 42 (6), 999–1012.

Andersen, E.B. (1970). "Asymptotic properties of conditional maximum likelihood estimation". Journal of the Royal Statistical Society, Series B 32 (2), 283–301.

Applebaum, E. (1982). "The estimation of the degree of oligopoly power". Journal of Econometrics 19, 287–299.

Athey, S., Haile, P.A. (2002). "Identification of standard auction models". Econometrica 70 (6), 2107–2140.

Bajari, P., Benkard, L. (2001). "Discrete choice models as structural models of demand: Some economic implications of common approaches". Working manuscript. Stanford Graduate School of Business.

Bajari, P., Benkard, L. (2005). "Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach". Journal of Political Economy 113 (6), 1239–1276.

Baker, J.B., Bresnahan, T.F. (1988). "Estimating the residual demand curve facing a single firm". International Journal of Industrial Organization 6 (3), 283–300.

Baron, D.P. (1989). "Design of regulatory mechanisms and institutions". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam.

Baron, D.P., Besanko, D. (1984). "Regulation, asymmetric information and auditing". RAND Journal of Economics 15 (4), 447–470.

Baron, D.P., Besanko, D. (1987). "Monitoring, moral hazard, asymmetric information, and risk-sharing in procurement contracting". Rand Journal of Economics 18 (4), 509–532.

Baron, D.P., Myerson, R. (1982). "Regulating a monopolist with unknown costs". Econometrica 50 (4), 911–930.

Becker, G.S. (1962). "Irrational behavior and economic theory". Journal of Political Economy 70 (1), 1–13.

Berry, S.T. (1992). "Estimation of a model of entry in the airline industry". Econometrica 60 (4), 889–917.

Berry, S.T. (1994). "Estimating discrete-choice models of product differentiation". RAND Journal of Economics 25 (2), 242–262.

Berry, S.T. (2001). "Estimating the pure hedonic choice model". Working manuscript. Yale Department of Economics.

Berry, S.T., Levinsohn, J., Pakes, A. (1995). "Automobile prices in market equilibrium". Econometrica 63 (4), 841–890.

Berry, S.T., Levinsohn, J., Pakes, A. (2004). "Estimating differentiated product demand systems from a combination of micro and macro data: The new car model". Journal of Political Economy 112 (1), 68–105.

Berry, S.T., Linton, O., Pakes, A. (2004). "Limit theorems for estimating the parameters of differentiated product demand systems". Review of Economic Studies 71, 613–654.

Berry, S.T., Reiss, P.C. (2003). "Empirical models of entry and exit". In: Porter, R.H., Armstrong, M. (Eds.), Handbook of Industrial Organization, vol. 3. North-Holland, Amsterdam. In press.

Besanko, D. (1984). "On the use of revenue requirements regulation under imperfect information". In: Crew, M.A. (Ed.), Analyzing the Impact of Regulatory Change in Public Utilities. Lexington Books, Lexington.

Blackorby, C., Primont, D., Russell, R.R. (1978). Duality, Separability and Functional Structure. North-Holland, Amsterdam.

Borenstein, S. (1992). "The evolution of US airline competition". Journal of Economic Perspectives 6 (2), 45–73.

Bresnahan, T.F. (1981). "Departures from marginal-cost pricing in the American automobile industry: Estimates for 1977–1978". Journal of Econometrics 11, 201–227.

Bresnahan, T.F. (1982). "The oligopoly solution concept is identified". Economics Letters 10 (1–2), 87–92.

Bresnahan, T.F. (1987). "Competition and collusion in the American automobile market: The 1955 price war". Journal of Industrial Economics 35 (4, June), 457–482.

Bresnahan, T.F. (1989). "Empirical methods for industries with market power". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam.

Bresnahan, T.F. (1997). "Comment". In: Bresnahan, T.F., Gordon, R. (Eds.), The Economics of New Goods. University of Chicago Press, Chicago.

Bresnahan, T.F., Reiss, P.C. (1985). "Dealer and manufacturer margins". RAND Journal of Economics 16 (2), 253–268.

Bresnahan, T.F., Reiss, P.C. (1991a). "Entry and competition in concentrated markets". Journal of Political Economy 99 (5), 977–1009.

Bresnahan, T.F., Reiss, P.C. (1991b). "Empirical models of discrete games". Journal of Econometrics 48 (1–2), 57–81.

Brueckner, J.K., Dryer, N.J., Spiller, P.T. (1992). "Fare determination in hub and spoke networks". RAND Journal of Economics 23 (3), 309–323.

Camerer, C. (1995). "Individual decision making". In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton University Press, Princeton.

Campo, S., Guerre, E., Perrigne, I.M., Vuong, Q. (2003). "Semiparametric estimation of first-price auctions with risk-averse bidders". Working manuscript. University of Southern California.

Christensen, L.R., Greene, W.H. (1976). "Economics of scale in US electric power generation". Journal of Political Economy 84 (4), 655–676.

Corts, K.S. (1999). "Conduct parameters and the measurement of market power". Journal of Econometrics 88 (2), 227–250.

Dalen, D.M., Gomez-Lobo, A. (1997). "Estimating cost functions in regulated industries characterized by asymmetric information". European Economic Review 41 (3–5), 935–942.

Davis P. (2000). "Demand models for market-level data". Working manuscript. MIT Sloan School.

Deaton, A., Muellbauer, J. (1980). Economic and Consumer Behavior. Cambridge University Press, Cambridge.

Dixit, A.K., Stiglitz, J.E. (1977). "Monopolistic competition and optimum product diversity". American Economic Review 67 (3), 297–308.

Dunne, T., Roberts, M.J., Samuelson, L. (1988). "Patterns of firm entry and exit in US manufacturing industries". RAND Journal of Economics 19 (4), 495–515.

Engel, E. (1857). "Die Productions-und Consumptionsverhältnisse des Königreichs Sächsen". In: Zeitschrift des Statischen Bureaus des Königlich Söchsischen Ministeriums des Inneren, Nos. 8 and 9.

Evans, D., Heckman, J.J. (1984). "A test of subadditivity of the cost function with an application to the Bell system". American Economic Review 74 (4), 615–623.
Gagnepain, P., Ivaldi, M. (2002). "Incentive regulatory policies: The case of public transit systems in France". RAND Journal of Economics 33 (4), 605–629.
Goldberg, P.K. (1995). "Product differentiation and oligopoly in international markets: The case of the US automobile industry". Econometrica 63 (4), 891–951.
Goldberger, A.S. (1991). A Course in Econometrics. Harvard University Press, Cambridge.
Gollop, F.M., Roberts, M.J. (1979). "Firm interdependence in oligopolistic markets". Journal of Econometrics 10 (3), 313–331.
Gorman, W.M. (1959). "Separable Utility and Aggregation". Econometrica 27 (3), 469–481.
Gorman, W.M. (1970). "Two-stage budgeting". In: Blackorby, C., Shorrocks, A. (Eds.), Separability and Aggregation. In: Collected Works of W.M. Gorman, vol. 1. Clarendon Press, Oxford.
Green, E.J., Porter, R.H. (1984). "Noncooperative collusion under imperfect price information". Econometrica 52 (1), 87–100.
Guerre, E., Perrigne, I.M., Vuong, Q. (2000). "Optimal nonparametric estimation of first-price auctions". Econometrica 68 (3), 525–574.
Haavelmo, T. (1944). "The probability approach in economics". Econometrica 12 (Suppl.), iii-vi and 1-115.
Haile, P.A., Tamer, E. (2003). "Inference with an incomplete model of English Auctions". Journal of Political Economy 111 (1), 1–51.
Hanemann, W.M. (1984). "Discrete/continuous models of consumer demand". Econometrica 52 (3), 541–562.
Härdle, W. (1990). Applied Nonparametric Regression. Cambridge University Press, London.
Härdle, W., Linton, O. (1994). "Applied nonparametric methods". In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.
Hausman, J. (1997). "Valuation of new goods under perfect and imperfect competition". In: Bresnahan, T.F., Gordon, R. (Eds.), The Economics of New Goods. University of Chicago Press, Chicago.
Hausman, J., Leonard, G., Zona, D. (1994). "Competitive analysis with differentiated products". Annales d'Economie et de Statistique 34 (0), 159–180.
Heckman, J.J. (1978). "Dummy endogenous variables in a simultaneous equation system". Econometrica 46 (4), 931–959.
Hendricks, K., Paarsch, H.J. (1995). "A survey of recent empirical work concerning auctions". Canadian Journal of Economics 28 (2), 403–426.
Hendricks, K., Porter, R.H. (1988). "An empirical study of an auction with asymmetric information". American Economic Review 78 (5), 865–883.
Hendricks, K., Porter, R.H. (2000). "Lectures on auctions: An empirical perspective". In: Armstrong, M., Porter, R.H. (Eds.), Handbook of Industrial Organization, vol. 3. North-Holland, Amsterdam. In press.
Hood, W.C., Koopmans, T.C. (1953). Studies in Econometric Method, Cowles Commission Monograph no. 14. John Wiley, New York.
Krasnokutskaya, E. (2002). "Identification and estimation of auction models under unobserved auction heterogeneity". Working manuscript. Yale Department of Economics.
Laffont, J.J. (1997). "Game theory and empirical economics: The case of auction data". European Economic Review 41 (1), 1–35.
Laffont, J.J., Tirole, J. (1986). "Using cost observation to regulate firms". Journal of Political Economy 94 (3), 614–641.
Laffont, J.J., Tirole, J. (1993). A Theory of Incentives in Procurement and Regulation. MIT Press, Cambridge.
Laffont, J.J., Vuong, Q. (1996). "Structural analysis of auction data". American Economic Review 36 (2), 414–420.
Laffont, J.J., Ossard, H., Vuong, Q. (1995). "Econometrics of first-price auctions". Econometrica 63 (4), 953–980.
Lau, L.J. (1982). "On identifying the degree of industry competitiveness from industry price and output data". Economics Letters 10 (1–2), 93–99.
Lerner, A. (1934). "The concept of monopoly and the measurement of monopoly power". Review of Economic Studies 1 (3), 157–175.

Li, T., Perrigne, I.M., Vuong, Q. (2002). "Structural estimation of the affiliated private value auction model". RAND Journal of Economics 33, 171–193.
Lindh, T. (1992). "The inconsistency of consistent conjectures". Journal of Economic Behavior and Organization 18 (1), 69–90.
Maddala, G.S. (1983). Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.
McAfee, R.P., McMillan, J. (1987). "Auctions and bidding". Journal of Economic Literature 25 (2), 699–738.
Milgrom, P.R., Weber, R.J. (1982). "A theory of auctions and competitive bidding". Econometrica 50 (5), 1089–1122.
Morrison, S.A., Winston, C. (1996). "The causes and consequences of airline fare wars". Brookings Papers on Economic Activity: Microeconomics, 85–123.
Nevo, A. (2000). "Mergers with differentiated products: The case of the ready-to-eat cereal industry". RAND Journal of Economics 31 (3), 395–421.
Ott, J. (1990). "Justice Dept. Investigates Carriers' Pricing Policies". Aviation Week and Space Technology 133 (3), 18–20.
Paarsch, H.J. (1992). "Deciding between the common and private values paradigms in empirical models of auctions". Journal of Econometrics 51 (1–2), 191–216.
Paarsch, H.J. (1997). "Deriving an estimate of the optimal reserve price: An application to British Columbia timber sales". Journal of Econometrics 78 (2), 333–357.
Petrin, A. (2002). "Quantifying the benefits of new products: The case of the minivan". Journal of Political Economy 110 (4), 705–729.
Phillips, A.W. (1958). "The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957". Economica 25 (100), 283–299.
Pinkse, J., Slade, M., Brett, C. (2002). "Spatial price competition: A semiparametric approach". Econometrica 70 (3), 1111–1155.
Pollak, R.A., Wales, T.J. (1992). Demand System Specification and Estimation. Oxford University Press, New York.
Porter, R.H. (1983). "A study of cartel stability: The Joint Executive Committee, 1880–1886". Bell Journal of Economics 14 (2), 301–314.
Quandt, R. (1988). The Econometrics of Disequilibrium. Basil Blackwell, Oxford.
Riordan, M.H. (1985). "Imperfect information and dynamic conjectural variations". RAND Journal of Economics 16 (1), 41–50.
Rosse, J.N. (1970). "Estimating cost function parameters without using cost data: Illustrated methodology". Econometrica 38 (2), 256–275.
Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
Sims, C.A. (1980). "Macroeconomics and reality". Econometrica 48 (1), 1–47.
Spiller, P.T., Favaro, E. (1984). "The effects of entry regulation on oligopolistic interaction: The Uruguayan Banking Sector". RAND Journal of Economics 15 (2), 244–254.
Stigler, G. (1964). "A theory of oligopoly". Journal of Political Economy 72 (1), 44–61.
Ulen, T.S. (1978). "Cartels and regulation". Unpublished PhD dissertation. Stanford University.
White, H. (1980). "Using least squares to approximate unknown regression functions". International Economic Review 21 (1), 149–170.
Windle, R. (1993). "Competition at 'Duopoly' Airline Hubs in the US". Transportation Journal 33 (2), 22–30.
Wolak, F.A. (1994). "An econometric analysis of the asymmetric information, regulator–utility interaction". Annales d'Economie et de Statistique 34 (0), 13–69.
Wolak, F.A. (2000). "An empirical analysis of the impact of hedge contracts on bidding behavior in a competitive market". International Economic Journal 14 (2), 1–40.
Wolak, F.A. (2003). "Identification and estimation of cost functions using observed bid data: An application to electricity". In: Detwatripont, M., Hansen, L.P., Turnovsky, S.J. (Eds.), Advances in Econometrics: Theory and Applications. In: Eighth World Congress, vol. 2. Cambridge University Press, Cambridge.

Chapter 65

MICROECONOMETRIC MODELS OF INVESTMENT AND EMPLOYMENT*

STEPHEN BOND

Nuffield College and Department of Economics, University of Oxford, UK

Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK
e-mail: steve.bond@nuffield.ox.ac.uk

JOHN VAN REENEN

Centre for Economic Performance, London School of Economics, UK

CEPR, UK

NBER, USA
e-mail: j.vanreenen@lse.ac.uk

Contents

* We thank Manuel Arellano, Nick Bloom, Jeffrey Campbell, Jason Cummins, Jan Eberly, Eduardo Engel, Dan Hamermesh, Lars Hansen, James Heckman, Tor Jakob Klette, Steve Nickell, John Pencavel and Frank Wolak for helpful comments and discussions. We thank the ESRC for financial support through the Centre for the Microeconomic Analysis of Fiscal Policy at the Institute for Fiscal Studies and the Centre for Economic Performance at the London School of Economics. All errors are our own.

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06065-5

Abstract

We survey recent microeconometric research on investment and employment that has used panel data on individual firms or plants. We focus on model specification and econometric estimation issues, but we also review some of the main empirical findings. We discuss advantages and limitations of microeconomic data in this context.

We briefly review the neoclassical theory of the demand for capital and labour, on which most of the econometric models of investment and employment that we consider are based. We pay particular attention to dynamic factor demand models, based on the assumption that there are costs of adjustment, which have played a prominent role especially in the microeconometric literature on investment. With adjustment costs, current choices depend on expectations of future conditions. We discuss the challenges that this raises for econometric model specification, and some of the solutions that have been adopted. We also discuss estimation issues that arise for dynamic factor demand equations in the context of micro panel data for firms or plants.

We then discuss a number of topics that have been the focus of recent microeconometric research on investment and employment. In particular, we review the literatures

on investment and financing constraints, relative price effects on investment and employment, investment and uncertainty, investment in research and development (R&D), elasticities of substitution and complementarity between technology, capital and skilled and unskilled labour, and recent work on models with non-convex adjustment costs.

Keywords

investment, employment, panel data

JEL classification: D92, G31, J23, D21, C50, O33

1. Introduction

This chapter surveys the application of econometric methods to study investment and employment decisions, using microeconomic data at the level of individual firms or plants. We discuss a range of models and issues that have been at the centre of microeconometric research on company factor demand behaviour over the last decade. We do not attempt to review the extensive econometric research on investment and employment that uses more aggregated data at the sectoral or macroeconomic levels. Chirinko (1993a) and Caballero (1999) provide surveys of recent work on aggregate investment, whilst Hamermesh (1993) provides a comprehensive survey of work on aggregate employment.

Microeconomic data offers several important advantages for the study of investment and employment behaviour. First, it allows us to eliminate the impact of aggregation over firms or plants when estimating a particular model. Second, there may be cross-sectional variation in explanatory variables that helps to identify parameters of interest. Perhaps more importantly, the availability of micro data allows the researcher to investigate heterogeneity in behaviour between different types of firms or plants that would simply not be possible with more aggregated data.

On the other hand, microeconomic data sources are often unrepresentative in their coverage (typically with a bias towards larger units), and severely limited in the type of information that they provide. These limitations have shaped the research questions that micro data sets have been widely used to address. A basic reason for estimating models of investment or labour demand is to quantify how the employment of capital or labour inputs responds to changes in factor prices – for example, the response of investment to interest rates or to an investment tax credit, or the response of employment to minimum wage legislation. These questions require elasticities of substitution to be estimated – serious investigation requires a flexible representation of substitution possibilities to be specified, and estimation requires detailed information on a range of factor inputs or cost shares, and the corresponding factor prices. Unfortunately many micro data sources, particularly those obtained from company accounts, provide only crude or partial measures of factor inputs or cost shares, and little or no information on the factor prices faced by individual firms or plants.

Largely for this reason, much of the microeconometric work has focused on estimating a single equation for company investment or company employment, rather than the more ambitious systems of interrelated factor demand or share equations that are often estimated using aggregate or industry level data.[1] Also, where this literature has investigated elasticities of substitution, considerable ingenuity has often been required, either to obtain measurable variation across firms in factor prices, or to specify models which can address some questions of interest without price information at all. Examples of the former include variation in the cost of capital due to a different mix between

[1] See Berndt (1991) for an excellent introduction to the interrelated factor demand literature.

tax-favoured and unfavoured types of investment, or variation in the effective impact of taxes due to asymmetries between the tax treatment of profits and losses; examples of the latter include work on technology-skill complementarity, which we discuss in Section 7.5 below.

Notwithstanding these limitations, there are a range of interesting questions that can be and have been addressed using microeconomic data. We are interested not only in how much investment or employment will respond to a change in factor prices, but also in how quickly. Once we recognise that complete adjustment does not occur immediately, the question of how current investment and employment decisions depend on expectations of future prices and demand conditions becomes important, and controlling for these unobserved expectations presents a particular challenge for econometric modelling. Indeed for some policy questions, such as the evaluation of temporary tax incentives, the characterisation of adjustment dynamics is of crucial importance. The recognition that aggregation can distort the underlying dynamic relationship has motivated the use of micro data to study adjustment processes for both capital and labour.[2] Moreover, as we discuss in Section 3, models have been developed that allow adjustment parameters to be estimated without necessarily restricting or identifying the long-run elasticities of substitution.

Another major use of micro data has been to test some of the simplifying assumptions that are made in specifying traditional factor demand models. Leading examples include the question of whether firms can finance their investment spending in 'perfect' capital markets, or whether they may face important financing constraints; and the question of whether firms determine employment along a labour demand schedule, or whether employment levels are subject to bargaining with workers. Micro data allows heterogeneity across firms – for example, between small and large firms, or between unionized and non-unionized plants – to be exploited in testing these specifications, and the importance of differences among sub-samples of firms or plants to be investigated. Again these questions can be addressed without fully specifying the nature of the production technology.

The interest in testing hypotheses about the nature of the adjustment process or the environment in which investment and employment decisions are taken has led to a focus on structural models of investment and employment dynamics, in which the optimal evolution of the firm's stock of capital or labour inputs is derived from some underlying theoretical model of the firm, and this is used to obtain an econometric model whose parameters reflect the firm's technology. As we discuss further in Section 3.6, reduced form dynamic models of investment or employment generally compound structural adjustment parameters with the process by which current expectations of future demand or prices are formed, which makes it difficult to draw firm conclusions about the nature of

[2] See Nickell (1978, 1986) for a discussion of aggregation biases in the context of dynamic investment and labour demand equations, and Blundell and Stoker (Chapter 68 in this volume) for a survey of aggregation issues more generally.

the adjustment process or the role of, say, financial variables in an investment equation. Nevertheless, this *relative* disadvantage of reduced form models compared to actual (as distinct from ideal) structural models should not be overstated. As we will emphasize, the most commonly used structural dynamic models of investment and employment are based on extreme simplifying assumptions, and are frequently rejected when subjected to mild empirical testing. Moreover, the recent empirical work on models with non-convex adjustment costs, whilst largely descriptive, has cast doubt on many of the structural models developed in the 1980s.

Two further limitations of commonly used microeconomic data sets should be noted at the outset. First, data on publicly traded companies or manufacturing plants already reflect a considerable degree of aggregation over investment decisions in many different types of equipment and structures, and employment decisions over many different types of workers. For large units, there may also be aggregation over inputs used in different lines of business, and for annual data there is the further question of aggregation over time. Although these micro data may be the most appropriate level of aggregation for investigating some questions – such as the relationship between investment and share prices, debt structures, or other aspects of corporate finance and corporate governance – we should recognise that they may still be too aggregated for identifying other parameters of interest. Whilst these concerns are also present in the study of household level data,[3] they are probably more severe in the case of data on large firms or large plants. Conversely, though, we should also recognise that if the object of interest is how aggregate investment or employment responds to some change in wages or prices, then the adjustment that we observe within existing firms or plants may be only one component of the aggregate response. In particular, adjustment which takes the form of the entry or exit of more or less capital intensive firms or plants is likely to be missed with commonly used data sets, and for the same reason there may be no simple relationship between aggregate adjustment dynamics and those observed at the micro level.[4] Although there is much we can learn from the study of investment and employment adjustment in microeconomic data sets, we should be cautious in extending these findings to address macroeconomic questions.

In this chapter we concentrate principally on the issues that have been the focus of recent econometric research using firm or plant level data. We present the state of the art as we perceive it and we do not hesitate to point out what we consider to be important weaknesses and omissions in this literature. Today's gaps are tomorrow's opportunities for important progress to be made in research. Whilst we recognise that important advances have been made in the specification and estimation of microeconometric investment and employment models in recent years, this is an area where the development of new data resources is presenting new challenges to traditional approaches, as well as exciting opportunities for richer structural models to be developed.

[3] See, for example, Bourguignon and Chiappori (1992) and the discussion in Blundell, MaCurdy and Meghir (Chapter 69 in this volume).

[4] See Caballero (1992) and Campbell and Fisher (2000) for further discussion.

The chapter is organised as follows. Section 2 sets out the basic neoclassical theory of factor demand, on which most of the econometric models we consider are based. Section 2.1 briefly reviews the static factor demand literature, whilst Section 2.2 introduces dynamic factor demand models based on the assumption that changing the level of factor inputs involves adjustment costs. Section 3 illustrates how this approach has been used to derive dynamic econometric investment equations. Sections 3.1–3.4 discuss alternative structural models based on strictly convex costs of adjustment, including the popular Q model of investment and the Euler equation approach; Section 3.5 discusses more recent work on models with non-convex adjustment costs; and Section 3.6 discusses the use of reduced form dynamic models in this context. Section 4 discusses some econometric issues that arise in the specification and estimation of these dynamic factor demand equations, particularly those of stochastic specification, and estimation using micro panel data for individual firms or plants. Section 5 discusses the sources of such data, and its limitations. Section 6 discusses some topics in the recent empirical literature on investment. Section 6.1 presents some basic empirical findings; Section 6.2 discusses the literature on testing for financing constraints; Section 6.3 discusses some recent research on taxes and investment; Section 6.4 discusses some recent work on uncertainty and investment; and Section 6.5 discusses microeconometric models of research and development (R&D) investment. Section 7 discusses some topics in the recent empirical literature on employment. Section 7.1 considers wage elasticities; Section 7.2 discusses models of employment determination with union bargaining; Section 7.3 discusses models of employment dynamics; Section 7.4 discusses whether adjustment costs are important for net changes in the level of employment or for gross flows of hiring and firing; and Section 7.5 discusses research on skill-biased technical change. Section 8 presents our summary of the main themes, omissions and opportunities for future research in this area.

2. Theoretical framework

We begin this chapter with a brief exposition of the neoclassical theory of factor demand. The model we consider is simplified in many respects. The firm's objective is to maximise the value of the equity owned by its shareholders, so that a host of interesting corporate control issues are assumed away. These shareholders are assumed to be risk neutral, so that the effects of risk on the firm's required rate of return are not considered. The firm issues no debt and pays no taxes, so that corporate financial policy is not considered. The firm operates in competitive markets and in a world characterised by symmetric information, so that strategic behaviour is not considered, and the firm is able to issue as much new equity as it chooses at an exogenously given required rate of return, determined by the riskless interest rate. Hence internal finance from retained profits and external finance from new share issues are perfect substitutes, and there is separability between the firm's real and financial decisions, as in the Modigliani–Miller

(1958, 1961) theorems. It is not our intention to suggest that these omitted considerations are unimportant or uninteresting. We will touch on some of these issues in later sections, but to do full justice to them would take us well beyond the scope of this chapter.

We distinguish between three types of factors of production: capital assets, which are owned by the firm and provide productive services for several time periods; labour inputs, which are hired by the firm each period; and current inputs, which are purchased by the firm but which are fully consumed in contributing to the current period's production. Capital assets, which may include both tangible assets like equipment and structures and intangible assets like knowledge and reputation, are durable, whilst labour and current inputs are not. However a more important distinction is whether the level of these inputs can be costlessly and immediately adjusted in response to new information. We first examine the static case which abstracts from any adjustment costs or delays.

2.1. Static factor demand

It is useful to briefly review static models of the firm's demand for capital and labour, in order to introduce some important concepts and to clarify how the structural dynamic models we consider later generalise this static framework. These static models also form the basis for most reduced form dynamic factor demand equations, which we discuss further in Section 3.6.

The basic factor demand model we consider can be characterised by the following optimisation problem for the firm

$$V_t(K_{t-1}) = \Big\{ \max_{I_t, L_t, M_t} \Pi_t(K_t, L_t, M_t, I_t) + \beta_{t+1} E_t\big[V_{t+1}(K_t)\big] \Big\}, \tag{2.1}$$

where V_t is the maximised value of the firm in period t, $\Pi_t(.)$ is the firm's net revenue function in period t, $K_t = (K_t^1, \ldots, K_t^N)$ is a vector of N types of capital inputs, $L_t = (L_t^1, \ldots, L_t^R)$ is a vector of R types of labour inputs, $M_t = (M_t^1, \ldots, M_t^S)$ is a vector of S types of current inputs, $I_t = (I_t^1, \ldots, I_t^N)$ is a vector of gross investments in each type of capital, $\beta_{t+1} = (1 + \rho_{t+1})^{-1}$ is the firm's discount factor, where ρ_{t+1} is the risk-free rate of interest between period t and period $t+1$, and $E_t[\,.\,]$ denotes the expected value conditional on information available in period t, where the expectation is taken over the distribution of future prices and interest rates.

The equation of motion for the capital inputs is

$$K_t^i = \big(1 - \delta^i\big) K_{t-1}^i + I_t^i \quad \text{for } i = 1, \ldots, N, \tag{2.2}$$

where δ^i is the rate of depreciation for capital of type i, assumed to be exogenous and fixed. Note that gross investment may be positive or negative, so that disinvestment is also assumed to be costless.

In the absence of any adjustment costs, the net revenue function may take the form

$$\Pi_t(K_t, L_t, M_t, I_t) = p_t F(K_t, L_t, M_t) - p_t^K I_t - w_t L_t - p_t^M M_t, \tag{2.3}$$

where $F(K_t, L_t, M_t)$ is the production function, p_t is the price of the firm's output, $p_t^K = (p_t^{K,1}, \ldots, p_t^{K,N})$ is a vector of prices for each type of capital goods, $w_t = (w_t^1, \ldots, w_t^R)$ is a vector of wage rates for each type of labour and $p_t^M = (p_t^{M,1}, \ldots, p_t^{M,S})$ is a vector of prices for each type of current inputs. Note that capital inputs are assumed to be purchased and owned by the firm, whilst labour inputs are assumed to be hired.[5]

The solution to the optimisation problem (2.1) subject to the constraints (2.2) can be characterised by the first-order conditions

$$-\left(\frac{\partial \Pi_t}{\partial I_t^i}\right) = \lambda_t^i \quad \text{for } i = 1, \ldots, N, \tag{2.4}$$

$$\lambda_t^i = \left(\frac{\partial \Pi_t}{\partial K_t^i}\right) + \left(1 - \delta^i\right)\beta_{t+1} E_t\left[\lambda_{t+1}^i\right] \quad \text{for } i = 1, \ldots, N, \tag{2.5}$$

$$\left(\frac{\partial \Pi_t}{\partial L_t^i}\right) = 0 \quad \text{for } i = 1, \ldots, R, \tag{2.6}$$

$$\left(\frac{\partial \Pi_t}{\partial M_t^i}\right) = 0 \quad \text{for } i = 1, \ldots, S, \tag{2.7}$$

where $\lambda_t^i = \frac{1}{1-\delta^i}\left(\frac{\partial V_t}{\partial K_{t-1}^i}\right)$ is the shadow value of inheriting one additional unit of capital of type i in period t. Equation (2.4) shows that the cost of acquiring additional units of each type of capital in period t will be equated to their shadow values. Equation (2.5) describes the evolution of these shadow values along the optimal path for the capital stocks, whilst Equations (2.6) and (2.7) are standard first-order conditions for the non-durable factors of production, equating the price of these inputs with their marginal revenue products (see Equation (2.3)).

For a price-taking firm, we have $-\left(\frac{\partial \Pi_t}{\partial I_t^i}\right) = p_t^{K,i}$ and $\left(\frac{\partial \Pi_t}{\partial K_t^i}\right) = p_t\left(\frac{\partial F}{\partial K_t^i}\right)$. Substituting these expressions into (2.4) and (2.5), respectively, combining these equations to eliminate λ_t^i and $E_t[\lambda_{t+1}^i]$ from (2.5) and rearranging yields

$$\left(\frac{\partial F}{\partial K_t^i}\right) = \frac{p_t^{K,i}}{p_t}\left(1 - \left(\frac{1-\delta^i}{1+\rho_{t+1}}\right)E_t\left[\frac{p_{t+1}^{K,i}}{p_t^{K,i}}\right]\right) = \frac{r_t^i}{p_t} \quad \text{for } i = 1, \ldots, N. \tag{2.8}$$

This shows that if the level of capital inputs can be freely adjusted, the marginal product of capital of type i will be equated in each period with the real user cost of capital $\left(\frac{r_t^i}{p_t}\right)$ for capital of type i [Jorgenson (1963)]. The user cost depends on the relative price of capital goods of type i, the firm's required rate of return, the depreciation rate for capital of type i, and the expected rate of change in the price of capital goods of type i. This is also the equilibrium price at which capital goods of type i could be rented for use in

[5] The model can of course accommodate capital inputs that are leased; these would be treated in a similar way as labour inputs are treated here.

period t in a competitive rental market, so the user cost is also known as the rental price of capital.[6]

2.1.1. Functional forms

To derive useful factor demand equations that can be estimated, we then need to parameterise the static production function $F(K_t, L_t, M_t)$. First, we consider the popular Constant Elasticity of Substitution (CES) functional form [Arrow et al. (1961)]. To illustrate this we use a two factor production structure in which there is a single capital good (K_t), a single labour input (L_t), and no other current inputs.[7] Assuming constant returns to scale, this production function has the form

$$Y_t = F(K_t, L_t) = \left(a_K K_t^{\rho} + a_L L_t^{\rho}\right)^{\frac{1}{\rho}}, \tag{2.9}$$

where $\rho = (\frac{\sigma-1}{\sigma})$ and σ is the elasticity of substitution between capital and labour. To ensure that the firm's value maximization problem has a solution in the absence of adjustment costs, we also assume that there is some degree of monopolistic competition and the firm faces a downward sloping demand curve for its output (Y_t) of the isoelastic form

$$p_t = B Y_t^{-\frac{1}{\eta^D}}, \tag{2.10}$$

where B is a demand shift parameter and $\eta^D > 1$ is the price elasticity of product demand. Under these conditions the demands for capital and labour have the convenient forms

$$K_t = a_K^{\sigma} Y_t \left(\frac{r_t}{p_t(1-\frac{1}{\eta^D})}\right)^{-\sigma}, \tag{2.11}$$

$$L_t = a_L^{\sigma} Y_t \left(\frac{w_t}{p_t(1-\frac{1}{\eta^D})}\right)^{-\sigma}, \tag{2.12}$$

giving the log-linear equations

$$\ln K_t = \sigma \ln a_K \left(1 - \frac{1}{\eta^D}\right) + \ln Y_t - \sigma \ln\left(\frac{r}{p}\right)_t, \tag{2.13}$$

$$\ln L_t = \sigma \ln a_L \left(1 - \frac{1}{\eta^D}\right) + \ln Y_t - \sigma \ln\left(\frac{w}{p}\right)_t, \tag{2.14}$$

[6] The static model in which the firm purchases durable capital inputs is formally equivalent to a static model in which capital inputs are leased at the rental price (r_t^i). A voluminous literature has considered how various tax structures impact on the user cost of capital. See, for example, Hall and Jorgenson (1967), King (1974), King and Fullerton (1984) and Jorgenson and Landau (1993).

[7] Or more realistically, we treat $F(K_t, L_t)$ as a production function for value-added rather than gross output.

which can be used as a basis for estimating the elasticity of substitution, or the responsiveness of factor intensities to changes in relative prices. Interpreted as expressions for the desired levels of factor inputs in the long run, these static factor demand equations form the basis for many reduced form models of investment and employment, as we discuss further in Section 3.6 below.

Next, we consider the case of more than two factors of production. It is simple to extend the basic CES production function to this case, but this imposes the unappealing restriction that the elasticity of substitution between all pairs of inputs is the same. To allow different patterns of substitution (or complementarity) between different factors requires the use of a more flexible functional form. In this context it is convenient to consider the dual of the firm's profit maximisation problem, in which the firm is assumed to minimise its costs taking the level of output as given. To illustrate this dual approach we will assume that the cost function can be written as a translog [Christensen, Jorgenson and Lau (1971, 1973)], which is a second-order approximation to an arbitrary functional form.[8]

For n variable factors of production $X_t = (X_{1t}, \ldots, X_{nt})$ – for example, the capital, labour and current inputs considered above – and their associated vector of factor prices $W_t = (W_{1t}, \ldots, W_{nt})$ – for example, the user costs, wage rates and input prices considered above – the translog cost function has the form

$$\begin{aligned}\ln C_t = {} & \ln \alpha_0 + \sum_{i=1}^{n} \alpha_i \ln W_{it} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} \ln W_{it} \ln W_{jt} \\ & + \alpha_Y \ln Y_t + \frac{1}{2} \gamma_{YY} (\ln Y_t)^2 + \sum_{i=1}^{n} \gamma_{iY} \ln W_{it} \ln Y_t \\ & + \phi_\tau t + \frac{1}{2} \phi_{\tau\tau} t^2 + \phi_{\tau Y} t \ln Y_t + \sum_{i=1}^{n} \phi_{\tau W_i} t \ln W_{it}, \end{aligned} \tag{2.15}$$

where the coefficients on time (t) represent technical change. The ϕ_τ and $\phi_{\tau\tau}$ coefficients reflect factor-neutral technical change, whilst the $\phi_{\tau W_i}$ coefficients reflect technical change that is biased towards factor i.[9] The cost minimising choices of input demands (X_{it}) are then conveniently expressed as log-linear cost share equations. From Shephard's (1953) Lemma we have

$$\frac{\partial \ln C_t}{\partial \ln W_{it}} = \frac{W_{it}}{C_t} \frac{\partial C_t}{\partial W_{it}} = \frac{W_{it} X_{it}}{C_t} = \alpha_i + \sum_{j=1}^{n} \gamma_{ij} \ln W_{jt} + \gamma_{iY} \ln Y_t + \phi_{\tau W_i} t, \tag{2.16}$$

[8] This has been a popular choice in the applied microeconometric literature. Alternative functional forms include the two-level CES and the Generalised Leontief [see Hamermesh (1993) or Berndt (1991) for a more extended discussion].

[9] See Chambers (1988) for an extensive discussion, and Section 7.5 below for a detailed treatment of skill-biased technical change.

where $\frac{W_{it}X_{it}}{C_t} = S_{it}$ is the share of factor i in total costs.

There are a series of economic restrictions that can be imposed on the system of equations in (2.16). In order to correspond to a well behaved production function, a cost function should be homogeneous of degree one in the vector of factor prices. That is, for a given level of output, total cost must increase proportionally when all prices increase proportionally. This implies the following relationships among the parameters

$$\sum_{i=1}^{n} \alpha_i = 1, \qquad \sum_{i=1}^{n} \gamma_{iY} = 0,$$
$$\sum_{j=1}^{n} \gamma_{ij} = \sum_{i=1}^{n} \gamma_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{n} \gamma_{ij} = 0. \tag{2.17}$$

The returns to scale (μ_t) can be computed as the inverse of the elasticity of costs with respect to output. Specifically,

$$\mu_t = \left(\frac{\partial \ln C_t}{\partial \ln Y_t}\right)^{-1}. \tag{2.18}$$

Therefore, constant returns to scale implies that (for all i)

$$\gamma_{iY} = 0 \quad \text{and} \quad \gamma_{YY} = 0, \tag{2.19}$$

and hence the output term drops out of the share equations (2.16) under this restriction.

Uzawa (1962) has shown that the Allen partial elasticities of substitution between two inputs i and j [Allen (1938)] can be computed from the cost formula

$$\sigma_{ijt} = \frac{C_t(\partial^2 C_t/\partial W_{it}\partial W_{jt})}{(\partial C_t/\partial W_{it})(\partial C_t/\partial W_{jt})}. \tag{2.20}$$

For the translog cost function we have

$$\sigma_{ijt} = \frac{\gamma_{ij} + S_{it}S_{jt}}{S_{it}S_{jt}}, \quad i \neq j, \tag{2.21}$$

and

$$\sigma_{iit} = \frac{\gamma_{ii} + S_{it}^2 - S_{it}}{S_{it}^2}. \tag{2.22}$$

The own and cross-price factor demand elasticities are given by $\epsilon_{ijt} = S_{jt}\sigma_{ijt}$ and are also easily calculated from the estimated parameters. We discuss examples of the implementation of this structure for the three-factor case with capital, skilled labour and unskilled labour in Section 7.5 below. Notice that the basic CES production function corresponds to the restriction that $\sigma_{iit} = \sigma_{ijt} = \sigma$ for all factors i and j and time periods t. The Cobb–Douglas production function [Cobb and Douglas (1928)] imposes the further restriction that $\sigma = 1$.

2.2. Dynamic factor demand

We now introduce adjustment costs. The basic motivation for introducing costs of adjustment is to account for the observation that adjustment in the level of factor inputs takes time to complete, or more specifically for the empirical failure of models which assume adjustment to be costless and immediate. The assumption that an activity is costly is a natural way for an economist to rationalise why more of it does not take place, but it is not the only possibility. Alternative models of investment have, for example, introduced exogenous delays (e.g. delivery lags) to explain investment dynamics, or considered limited substitution possibilities between installed capital and variable factors of production.[10]

We will continue to assume that current inputs are variable factors, in the sense that the level of these inputs can be varied immediately and without paying any adjustment costs. We will assume that capital inputs are quasi-fixed factors, in the sense that variations in their level impose costs of adjustment on to the firm, which will tend to delay and may possibly prevent their adjustment in response to new information. Thus at any point in time the level of capital inputs may differ from those which satisfy the static first-order conditions (like those illustrated for a price-taking firm in Equation (2.8)); and if there are adjustment costs associated with, for example, replacement investment, then the steady-state level of capital inputs may also differ from the solution to the static problem with costless adjustment. We will also allow labour inputs to be subject to adjustment costs. Define $H_t = (H_t^1, \ldots, H_t^R)$ as a vector of gross hiring in each type of labour, and the equation of motion for labour as

$$L_t^i = \big(1 - \gamma^i\big)L_{t-1}^i + H_t^i \quad \text{for } i = 1, \ldots, R, \tag{2.23}$$

where γ^i is the quit rate for labour of type i, also assumed to be exogenous and fixed.

The value maximisation problem analogous to (2.1) is now

$$\begin{aligned} &V_t(K_{t-1}, L_{t-1}) \\ &\quad = \Big\{ \max_{I_t, H_t, M_t} \Pi_t(K_t, L_t, M_t, I_t, H_t) + \beta_{t+1} E_t\big[V_{t+1}(K_t, L_t)\big] \Big\} \end{aligned} \tag{2.24}$$

where the dependence of net revenue on gross investment and gross hiring now reflects the presence of adjustment costs. The net revenue function is now specified as

$$\begin{aligned} &\Pi_t(K_t, L_t, M_t, I_t, H_t) \\ &\quad = p_t\big[F(K_t, L_t, M_t) - G(I_t, H_t, K_t, L_t)\big] - p_t^K I_t - w_t L_t - p_t^M M_t, \end{aligned} \tag{2.25}$$

where $G(I_t, H_t, K_t, L_t)$ is the adjustment cost function, with adjustment costs assumed to take the form of foregone production and initially assumed to be strictly convex in gross investment and gross hiring.

[10] See Jorgenson (1971) and Nickell (1978) for comprehensive accounts of these approaches in the investment context.

Given this specification, the solution to the firm's value maximisation problem continues to be characterised by first-order conditions (2.4), (2.5) and (2.7). The new first-order conditions for hiring and labour are

$$-\left(\frac{\partial \Pi_t}{\partial H_t^i}\right) = \mu_t^i \quad \text{for } i = 1, \ldots, R, \tag{2.26}$$

$$\mu_t^i = \left(\frac{\partial \Pi_t}{\partial L_t^i}\right) + \left(1 - \gamma^i\right)\beta_{t+1} E_t\left[\mu_{t+1}^i\right] \quad \text{for } i = 1, \ldots, R, \tag{2.27}$$

where $\mu_t^i = \frac{1}{1-\gamma^i}\left(\frac{\partial V_t}{\partial L_{t-1}^i}\right)$ is the shadow value of inheriting one additional unit of labour of type i in period t.

We have assumed that these costs of adjustment are strictly convex and differentiable, which will tend to smooth the adjustment of quasi-fixed factors to new information, since a series of small adjustments is assumed to be cheaper than a single large change in the level of these inputs. This limitation on the flexibility of factor inputs also rationalises forward-looking factor demand behaviour: since a change in the capital stock, for example, will be costly to reverse subsequently, the response of investment to a given change in the cost of capital will be different depending on whether that change is expected to be temporary or permanent. Hence these models predict that expectations of future demand and prices will be important determinants of current investment and employment decisions, which presents a particular challenge for econometric modelling. In the next section we discuss several approaches to this problem that have been used in the investment literature, beginning with those that can be obtained as special cases of the general factor demand model presented here. Specifications of dynamic labour demand models are discussed more briefly in Section 7.

Whilst most structural models of investment and employment dynamics that have been widely used in modelling factor demands at the firm level in the last twenty years have been based on the assumption of strictly convex adjustment costs, it should be noted that this specification was introduced into the literature principally as a matter of analytical convenience. More recent models have considered the implications of non-convex costs of adjustment, either by assuming (partial) irreversibility of current investment decisions, or by introducing a fixed cost component in the specification of adjustment costs. These models, which predict large but infrequent adjustments, will be considered further in Section 3.5 below.

3. Dynamic investment models

Most of the structural models of investment used in empirical analysis of firm level data can be obtained as special cases of the general factor demand model outlined in the previous section. Most of these models assume a single (homogeneous) capital input, and treat this as the only quasi-fixed factor used by the firm. The most popular of these models has been the Q model, which requires assumptions under which the unobserved

shadow value of capital is simply related to the observed market-to-book or average q ratio. Specialising the assumption of strictly convex adjustment costs to a symmetric, quadratic functional form then yields a convenient, linear equation based on (2.4), relating investment to observed average q. Although still widely used, dissatisfaction with the empirical performance of the Q model has led to interest in less restrictive implementations of the basic adjustment costs model, such as the approach proposed by Abel and Blanchard (1986) and the Euler equation approach, introduced into the investment literature by Abel (1980). We consider each of these models in turn.

3.1. The Q model

To illustrate the Q model we assume that the firm's only quasi-fixed input is a single homogeneous capital good. Most realistically, this can be thought of as a fixed coefficients aggregate of the different capital goods used by the firm.[11] Provided there are no adjustment costs associated with labour and current inputs, the model can straightforwardly allow for many types of these inputs. We consider the case of multiple quasi-fixed factors in Section 3.4 below.

Based on the net revenue equation (2.25) and the assumption of perfectly competitive markets, we then obtain

$$\left(\frac{\partial \Pi_t}{\partial I_t}\right) = -p_t\left(\frac{\partial G}{\partial I_t}\right) - p_t^K, \tag{3.1}$$

which substituted into (2.4) yields

$$\left(\frac{\partial G}{\partial I_t}\right) = \left(\frac{\lambda_t}{p_t^K} - 1\right)\frac{p_t^K}{p_t} = (q_t - 1)\frac{p_t^K}{p_t}. \tag{3.2}$$

Moreover, solving Equation (2.5) forward by repeated substitution yields

$$\lambda_t = E_t\left[\sum_{s=0}^{\infty}(1-\delta)^s \beta_{t+s}\left(\frac{\partial \Pi_{t+s}}{\partial K_{t+s}}\right)\right], \tag{3.3}$$

where β_{t+s} is the discount factor that discounts period $t+s$ revenues back to period t.

To interpret these expressions, notice that the shadow value of an additional unit of capital (λ_t) is a forward-looking measure of current and expected future values of the marginal revenue product of capital, where the discounting reflects the diminution of each current unit of capital over time through depreciation, as well as the standard compensation for delay. In the static factor demand model, the optimal capital stock was characterised by $\lambda_t = p_t^K$, or by $q_t = \frac{\lambda_t}{p_t^K} = 1$, where this ratio of shadow value to purchase cost is known as *marginal q*. With strictly convex costs of adjustment, marginal adjustment costs ($\frac{\partial G}{\partial I_t}$) are an increasing function of current gross investment, so

[11] For example, the firm's technology may be such that it must always combine 2 units of equipment with 1 unit of structures, but it can substitute between this capital aggregate and other labour and current inputs.

Equation (3.2) shows that investment is an increasing function of the deviation between the actual value of marginal q and this desired value in the absence of adjustment costs. Moreover, we have the striking result that all influences of expected future profitability on current investment are summarised in marginal q, through the shadow value of capital.

To obtain an empirical investment model we require an explicit form for marginal adjustment costs, and a way of measuring marginal q. Primarily for convenience, most implementations of the Q model have assumed that the cost of adjusting the capital stock is symmetric and quadratic about some 'normal' rate of investment, which may or may not be related to the rate of depreciation. More fundamentally, the basic Q model requires the adjustment cost function $G(I_t, K_t)$ to be homogeneous of degree one in (I_t, K_t), consistent with constant returns to scale. One popular functional form that has these properties, suggested by Summers (1981), is

$$G(I_t, K_t) = \frac{b}{2}\left[\left(\frac{I}{K}\right)_t - a\right]^2 K_t, \tag{3.4}$$

where the parameter b reflects the importance of adjustment costs. Using this specification in (3.2) gives the linear model

$$\left(\frac{I}{K}\right)_t = a + \frac{1}{b}\left[(q_t - 1)\frac{p_t^K}{p_t}\right]. \tag{3.5}$$

The distinctive feature of the Q model is the equality between marginal q and average q established by Hayashi (1982). The basic requirement is that the net revenue function $\Pi_t(K_t, L_t, M_t, I_t)$ is homogeneous of degree one, sufficient conditions for which are that both the production function and the adjustment cost function display constant returns to scale, and the firm is a price taker in all markets.[12] In this case we can combine Equations (2.4) and (2.5) to obtain

$$\lambda_t(K_t - I_t) = \left(\frac{\partial \Pi_t}{\partial I_t}\right) I_t + \left(\frac{\partial \Pi_t}{\partial K_t}\right) K_t + \beta_{t+1} E_t\left[(1-\delta)\lambda_{t+1} K_t\right] \tag{3.6}$$

or

$$(1-\delta)\lambda_t K_{t-1} = \Pi_t(K_t, L_t, M_t, I_t) + \beta_{t+1} E_t\left[(1-\delta)\lambda_{t+1} K_t\right], \tag{3.7}$$

since $(\frac{\partial \Pi_t}{\partial L_t}) = (\frac{\partial \Pi_t}{\partial M_t}) = 0$ for the variable factors of production. Solving forward by repeated substitution gives

$$(1-\delta)\lambda_t K_{t-1} = E_t\left[\sum_{s=0}^{\infty} \beta_{t+s} \Pi_{t+s}(K_{t+s}, L_{t+s}, M_{t+s}, I_{t+s})\right] = V_t, \tag{3.8}$$

[12] The presence of strictly convex adjustment costs in this model ensures that the value maximization problem has a solution, even with perfect competition and constant returns.

where V_t is again the maximised value of the firm. Thus we have

$$\lambda_t = \frac{V_t}{(1-\delta)K_{t-1}} \quad \text{or} \quad q_t = \frac{V_t}{(1-\delta)p_t^K K_{t-1}} \tag{3.9}$$

so that marginal q is equal to the ratio of the maximised value of the firm in period t to the replacement cost value in period t of the capital stock that the firm inherits from the previous period. This ratio, known as average q or Tobin's q [Brainard and Tobin (1968), Tobin (1969)], can in principle be measured.[13] The usual implementation further requires that share prices are not affected by bubbles or fads, so that the 'fundamental' value of the firm given in (3.8) can be measured by its stock market valuation.[14]

Substituting average q for marginal q in (3.5) then gives the basic Q investment equation as

$$\begin{aligned}\left(\frac{I}{K}\right)_t &= a + \frac{1}{b}\left[\left(\frac{V_t}{(1-\delta)p_t^K K_{t-1}} - 1\right)\frac{p_t^K}{p_t}\right] \\ &= a + \frac{1}{b}Q_t.\end{aligned} \tag{3.10}$$

Notice that if share prices do correctly reflect fundamentals then the structure of the Q model implies that all relevant expectations of future profitability are summarised by the firm's stock market valuation, and the prediction that Q_t defined in (3.10) should be a sufficient statistic for investment.

One further point to notice is that the Q model identifies the parameters (a, b) of the adjustment cost function (3.4). These parameters are identified without requiring any functional form for the gross production function to be specified, given the assumptions of perfect competition and constant returns to scale. This may be an advantage or a disadvantage, depending on the context. If the objective is to quantify the importance of adjustment costs, or to test this specification, then this robustness to different functional forms for the production function may be an advantage. On the other hand, if the objective is to estimate the response of investment to some change in tax rates or other component of the user cost of capital, then it is not sufficient to know the parameters of the adjustment cost function. Simulating the effects of a tax change on investment would require additional information about the elasticity of substitution between capital and other factors of production that is not identified by estimation of the Q model alone.[15]

[13] Abel and Eberly (1994) show more generally that average q is proportional to marginal q if the net revenue function is homogeneous of degree k. The Hayashi (1982) equality result is the special case with $k = 1$.

[14] The model extends straightforwardly to incorporate exogenous debt policies, in which case the numerator of average q becomes the maximised value of the firm's capital assets, and the firm's equity market valuation has to be adjusted by an estimate of the firm's outstanding debt. The model can also be extended for various forms of taxation. See, for example, Summers (1981) and Hayashi (1982).

[15] See Summers (1981) and Salinger and Summers (1983) for examples of tax simulation in the context of the Q model.

We have highlighted the restrictive structure required to equate marginal q and average q, and the particular structure of adjustment costs needed to obtain a linear relationship between investment and marginal q. The advantages of the Q model compared to reduced form models are that the influence of expectations on current investment decisions is explicitly modelled, and that the parameters identified by estimating Equation (3.10) are technological parameters of the adjustment cost function, which should be invariant to structural breaks in the underlying processes generating prices and interest rates. These are important advantages in the literature which tests the null hypothesis of perfect capital markets against an alternative in which financing constraints are important, as we discuss in Section 6.2 below.

Nevertheless there is no shortage of reasons why the Q model may be seriously misspecified. Adjustment costs may not be well described by the symmetric, quadratic functional form that is commonly imposed – thus the relationship between investment rates and Q_t may be non-linear and asymmetric even *within* the convex adjustment costs framework. Perfect competition and constant returns to scale may not be adequate assumptions, in which case average q ceases to be a sufficient statistic for the influence of expectations, although marginal q may still be.[16] Stock market valuations may differ from fundamental values without necessarily violating weaker forms of the efficient markets hypothesis [Summers (1986)] – for example, share prices may be affected by rational bubbles [Blanchard and Watson (1982)] or liquidity traders [Campbell and Kyle (1993)]. This would introduce severe measurement error problems if the average q ratio is constructed using share price data, and could undermine identification of the adjustment cost parameters.[17] Note also that for share prices to measure marginal q appropriately, the stock market must have the same expectations as the firm, in particular about the future path of the firm's capital stock. More mundane measurement error issues may also be important – the capital stock and debt measures that can be constructed from company accounts data are likely to be subject to substantial errors, and accurately measuring the (tax-adjusted) prices of capital goods relevant for individual firms is generally not possible with public data sets. It may be inappropriate to treat a single capital aggregate as the only quasi-fixed factor of production, and it may be necessary to distinguish between installed capital and new capital when specifying the substitution possibilities between capital and other factors of production. Firms may operate in imperfect capital markets, and the objectives of their managers may not always coincide with shareholder value maximisation.

Thus it may not be completely surprising that the empirical performance of the Q model has generally been disappointing. Estimates of Equation (3.10) have generally yielded extremely low values for the coefficient $(\frac{1}{b})$, suggesting extraordinarily high marginal costs of adjustment and implausibly slow adjustment of the actual capital stock. The prediction that Q_t is a sufficient statistic for investment has generally

[16] See Hayashi (1982).

[17] See Erickson and Whited (2000) and Bond and Cummins (2001) for further discussion.

been rejected in empirical tests, and the explanatory power of the Q_t variable is often found to be very weak when other variables such as sales and cash flow are added to the econometric model. Most attempts to control for measurement error in Q_t have been unsuccessful in reversing these basic results, although Cummins, Hassett and Hubbard (1994) find that the Q model provides a more satisfactory description of investment behaviour in periods when the variation in measured Q_t is dominated by tax changes. Recently Erickson and Whited (2000) and Bond and Cummins (2001) have also reported more favourable results when accounting for forms of measurement error suggested by the possibility of bubbles or fads in stock market valuations. These findings are consistent with the interpretation that variation in Q_t as normally measured is dominated by uninformative 'noise' in share prices, although as we have emphasised there are many other reasons why the Q model may be mis-specified.

3.2. The Abel and Blanchard model

If the concern is that the conditions required to equate marginal q and average q may not hold, or if measures of average q based on stock market valuations are suspect, then an alternative approach is to attempt to measure marginal q itself, and estimate Equation (3.5) directly.

Abel and Blanchard (1986) suggested constructing an estimate of the shadow value of capital using an auxiliary econometric model, based on Equation (3.3). This procedure requires a specification for the marginal revenue product of capital in terms of observable variables, and a forecasting model for these variables. Notice that this forecasting model does not need to yield accurate predictions, but rather needs to mimic the expectations of future marginal revenue products on which the firm's investment decisions are based. How well this can be done using the information available to the econometrician is not entirely clear. Given a set of forecasts of the future marginal revenue products of capital, these are discounted back to the current period to yield an estimate of λ_t. This is then used to construct an estimate of marginal q, which can be used in place of the average q ratio to estimate the investment equation.

This procedure avoids the use of share price data, and can in principle be used to relax the assumptions of perfect competition and constant returns to scale if a suitable form for the marginal revenue product of capital is specified. These assumptions are replaced by a specification for the marginal revenue product, and the need to specify an auxiliary forecasting model. The linear specification of the investment model still relies heavily on the assumption of symmetric, quadratic costs of adjustment.

3.3. The Euler equation

The Euler equation approach introduced by Abel (1980) can also relax the linear homogeneity of the net revenue function and avoid the use of share price data. Perhaps more importantly, this approach avoids the need to parameterise the expectations-formation

process. This is achieved by using the first-order condition for investment (2.4) to eliminate the shadow value of capital from the Euler equation (2.5), and then estimating the Euler equation itself rather than a model based on (2.4).

For the case of a single capital input this gives the expression

$$-\left(\frac{\partial \Pi_t}{\partial I_t}\right) = -(1-\delta)\beta_{t+1}E_t\left[\left(\frac{\partial \Pi_{t+1}}{\partial I_{t+1}}\right)\right] + \left(\frac{\partial \Pi_t}{\partial K_t}\right). \tag{3.11}$$

Using the net revenue function (2.25) and assuming perfectly competitive markets then gives

$$\left(\frac{\partial G}{\partial I_t}\right) = E_t\left[\psi_{t+1}\left(\frac{\partial G}{\partial I_{t+1}}\right)\right] + \left[\left(\frac{\partial F}{\partial K_t}\right) - \left(\frac{\partial G}{\partial K_t}\right) - \left(\frac{r}{p}\right)_t\right], \tag{3.12}$$

where $\psi_{t+1} = (\frac{1-\delta}{1+\rho_{t+1}})\frac{p_{t+1}}{p_t}$ is a real discount factor and $(\frac{r}{p})_t$ is the user cost of capital, as defined in Equation (2.8).

Comparing Equations (3.12) and (3.2) shows that the two terms on the right-hand side of (3.12) contain essentially the same information as marginal q. In particular, given the current difference between the marginal product of capital and the user cost, all relevant information about expected future profitability is here summarised by the one-step ahead forecast of discounted marginal adjustment costs.

Using the adjustment cost function (3.4) yields

$$\begin{aligned}\left(\frac{I}{K}\right)_t &= a\big(1 - E_t[\psi_{t+1}]\big) + E_t\left[\psi_{t+1}\left(\frac{I}{K}\right)_{t+1}\right] \\ &\quad + \frac{1}{b}\left[\left(\frac{\partial F}{\partial K_t}\right) - \left(\frac{\partial G}{\partial K_t}\right) - \left(\frac{r}{p}\right)_t\right].\end{aligned} \tag{3.13}$$

To implement this model, the one-step ahead expected values can be replaced by the realised values of these variables in period $t+1$, which introduces forecast errors that will be orthogonal to information available in period t under the assumption of rational expectations. Assuming constant returns to scale, the marginal product of capital can be substituted without assuming a parametric form for the production function, as in Bond and Meghir (1994). Alternatively the model can be implemented by assuming some form for the production function, as in Abel (1980). Whilst this is more restrictive, it shows that in principle substitution parameters can be identified from the Euler equation, which is not the case for the Q model. The Euler equation model can also be extended to allow for imperfectly competitive product markets and/or for decreasing returns to scale.

3.4. Multiple quasi-fixed factors

All the models we have considered thus far in this section have treated capital as a single quasi-fixed input, and assumed that all other inputs can be adjusted costlessly. The more general model outlined in Section 2.2 can be used to show how these models are affected

by the presence of more than one quasi-fixed factor. To illustrate this, we consider the case in which the firm can substitute between two types of capital (e.g. equipment and structures), both of which are subject to adjustment costs. The implications of treating labour as a quasi-fixed factor of production are essentially similar.

Combining Equations (2.4) and (2.5) as we did to obtain Equation (3.6), assuming $\Pi_t(K_t^1, K_t^2, L_t, M_t, I_t^1, I_t^2)$ is again homogeneous of degree one, and summing across the two types of capital yields

$$\sum_{i=1}^{2}(1-\delta^i)\lambda_t^i K_{t-1}^i = \Pi_t + \beta_{t+1}E_t\left[\sum_{i=1}^{2}(1-\delta^i)\lambda_{t+1}^i K_t^i\right] = V_t. \tag{3.14}$$

Thus marginal q for the first type of capital can be expressed as

$$q_t^1 = \frac{\lambda_t^1}{p_t^{K,1}} = \frac{V_t}{(1-\delta^1)p_t^{K,1}K_{t-1}^1} + \frac{1}{p_t^{K,1}}\left(\frac{\partial \Pi_t}{\partial I_t^2}\right)\left(\frac{1-\delta^2}{1-\delta^1}\right)\left(\frac{K_{t-1}^2}{K_{t-1}^1}\right), \tag{3.15}$$

and similarly for q_t^2. Assuming for simplicity that the adjustment cost function is additively separable in the two types of investment, such as

$$G(I_t^1, I_t^2, K_t^1, K_t^2) = \frac{b_1}{2}\left[\left(\frac{I_t^1}{K_t^1}\right) - a_1\right]^2 K_t^1 + \frac{b_2}{2}\left[\left(\frac{I_t^2}{K_t^2}\right) - a_2\right]^2 K_t^2, \tag{3.16}$$

we obtain a model for investment in the first type of capital as

$$\begin{aligned}\left(\frac{I_t^1}{K_t^1}\right) = a_1 &+ \frac{1}{b_1}\left[\left(\frac{V_t}{(1-\delta^1)p_t^{K,1}K_{t-1}^1} - 1\right)\frac{p_t^{K,1}}{p_t}\right] \\ &- \frac{b_2}{b_1}\left(\frac{1-\delta^2}{1-\delta^1}\right)\left(\frac{I_t^2}{K_t^2}\right)\left(\frac{K_{t-1}^2}{K_{t-1}^1}\right) + \frac{b_2 a_2}{b_1}\left(\frac{1-\delta^2}{1-\delta^1}\right)\left(\frac{K_{t-1}^2}{K_{t-1}^1}\right) \\ &- \frac{1}{b_1}\left(\frac{1-\delta^2}{1-\delta^1}\right)\left(\frac{p_t^{K,2}}{p_t}\right)\left(\frac{K_{t-1}^2}{K_{t-1}^1}\right),\end{aligned} \tag{3.17}$$

and similarly for investment in the second type of capital.

This shows that the basic Q model is mis-specified when there is more than one quasi-fixed factor. Whilst this system of equations could in principle be estimated, the literature has looked instead for restrictions on the form of adjustment costs that allow a single equation for total investment to be estimated. In particular, Hayashi and Inoue (1991) obtain conditions under which the structure of the basic Q model is preserved, with the aggregate capital stock measure being constructed as a Divisia index of the individual capital stocks, rather than as their sum.[18]

Both the Abel and Blanchard approach and the Euler equation approach extend more straightforwardly to the case of additively separable adjustment costs. An expression

[18] See also Galeotti and Schiantarelli (1991) and Chirinko (1993b).

analogous to (3.3) holds for each type of capital. This can be used to obtain an estimate of marginal q for each type of capital, provided one is willing to specify the marginal revenue product for each type of capital. This will again yield a system of equations for each type of investment. For the Euler equation case, expressions analogous to (3.11) and (3.12) hold for each type of capital.

These approaches can also accommodate the case of interrelated adjustment costs, provided one is willing to specify the form of the adjustment cost function. See, for example, Shapiro (1986) for a system of Euler equations with interrelated adjustment costs.

3.5. Non-convex adjustment costs

Several considerations have motivated researchers to consider models with non-convex costs of adjustment. As noted earlier, the assumption of strictly convex adjustment costs was introduced primarily for analytical convenience. Descriptive evidence on the time series behaviour of investment in very disaggregated data sets has questioned the empirical validity of this simplifying assumption. Analytical techniques that characterise optimal investment behaviour in the presence of non-convex adjustment costs have become more familiar to economists. And the empirical performance of the structural dynamic models based on strictly convex adjustment costs has been sufficiently problematic to encourage investigation of alternatives.

Census of production data provides evidence at the level of manufacturing establishments (plants, or related groups of plants), which is much more disaggregated than accounting data on companies. Doms and Dunne (1998) use data from the Longitudinal Research Database (LRD), covering 12,000 US manufacturing plants over the period 1972–1989, to illustrate several aspects of the 'lumpy' adjustment of capital. For example, more than half of all plants experience a year in which the capital stock increases by over 35%; the two largest investment 'spikes' are often observed in consecutive years; and the incidence of these investment spikes is highly correlated with the time series of aggregate investment for these establishments. Anti Nilsen and Schiantarelli (2003) report similar findings using plant level data for Norway. Their results also show that around 30% of Norwegian plants have zero investment in an average year, although this proportion falls to only 6% if they focus on main production plants, or aggregate over multiple plants belonging to the same company. Although aggregation over plants to the company level tends to smooth some of the discreteness observed for individual plants, Abel and Eberly (1996) find that the distribution of investment rates is positively skewed in a large sample of publicly traded US companies from the Compustat database, and the two highest investment rates are observed in consecutive years for about half the firms in their sample.

This evidence of infrequent and lumpy adjustment can be explained by the presence of adjustment cost components that are not strictly convex in the level of investment.[19]

[19] The same evidence could also be explained by indivisibilities, ruling out small purchases of capital. See Nickell (1978) for further discussion.

Complete irreversibility occurs when gross investment is restricted to be non-negative. As Abel and Eberly (1996) report that some disinvestment (sales of capital) occurs in about 50% of their observations, this assumption appears too extreme to characterise the investment behaviour of most companies.[20] Partial irreversibility occurs when used capital goods can only be sold for less than their true replacement cost value, for example as a result of adverse selection in second-hand capital goods markets [cf. Akerlof (1970)]. This wedge between the purchase price and the sale price of capital goods introduces a piecewise linear cost of adjustment, that is kinked at the point of zero gross investment. This kink is sufficient to explain why zero investment may occur even though the firm is not at its desired capital stock, but does not explain why observed adjustment should be lumpy. The latter can be explained by assuming a fixed cost component to adjustment costs, that is independent of the level of investment undertaken, although the size of the fixed cost may depend on whether the adjustment is upwards (positive investment) or downwards (disinvestment).

Theoretical models of investment have long since considered behaviour in the presence of irreversibility and linear adjustment costs [see, for example, Arrow (1968) and Nickell (1978)]. However it is only more recently, and notably through the work of Dixit and Pindyck (1994) and Bertola and Caballero (1994), that these models have begun to have a major impact on the empirical literature.

Abel and Eberly (1994, 1996) have extended the Q model of investment to obtain a structural model of investment dynamics in the presence of non-convex adjustment costs. As in the traditional Q model, they assume constant returns to scale and perfect competition in all markets. Their adjustment cost function has the form

$$G(I_t, K_t) = \begin{cases} a^+ K_t + b^+ I_t + c^+ (\frac{I}{K})_t^2 K_t & \text{if } I_t > 0, \\ a^- K_t + b^- I_t + c^- (\frac{I}{K})_t^2 K_t & \text{if } I_t < 0, \end{cases} \tag{3.18}$$

where aK_t denotes a fixed cost of adjustment that is paid if investment is non-zero, bI_t denotes a linear adjustment cost, $c(\frac{I}{K})_t^2 K_t$ denotes a strictly convex adjustment cost, and each of the parameters is allowed to take different values depending on the sign of gross investment. In this case the investment rate can be characterised as

$$\left(\frac{I}{K}\right)_t = \begin{cases} \frac{1}{c^+}(\lambda_t - p_t^K - b^+) & \text{if } \lambda_t > \bar{\lambda}_t(a^+, p_t^K + b^+), \\ 0 & \text{if } \underline{\lambda}_t(a^-, p_t^K - b^-) \leqslant \lambda_t \leqslant \bar{\lambda}_t(a^+, p_t^K + b^+), \\ \frac{1}{c^-}(\lambda_t - p_t^K - b^-) & \text{if } \lambda_t < \underline{\lambda}_t(a^-, p_t^K - b^-), \end{cases} \tag{3.19}$$

where λ_t is the shadow value of capital, $\bar{\lambda}_t$ is an upper threshold value below which the firm does not find it worthwhile to undertake positive gross investment, and $\underline{\lambda}_t$ is a lower threshold above which the firm does not find it worthwhile to undertake disinvestment.

[20] This is not to suggest that models with complete irreversibility are not useful in characterising some investment decisions, for example the decision to develop an offshore oil field. See, for example, Pesaran (1990) and Favero and Pesaran (1994).

Moreover, since this adjustment cost function is homogeneous of degree one in (I_t, K_t), the model can be implemented by equating marginal q with average q in the standard way. This model suggests a monotonic but non-linear relationship between investment and average q, with a region of inactivity where gross investment is zero between the two threshold values.

Caballero and Leahy (1996) have criticised the Abel and Eberly formulation. They show that the assumptions of perfect competition and constant returns to scale take on a more crucial role in the presence of fixed costs of adjustment. Not only is linear homogeneity of the net revenue function required to equate marginal and average q, but in the presence of fixed adjustment costs this restriction is also needed to obtain a monotonic relationship between investment and *marginal* q. If the net revenue function is concave, they show that marginal q becomes a non-monotonic function of investment and the inverse function, investment as a function of marginal q, does not exist. Thus the model described by (3.19) would be fundamentally mis-specified, however one measures the shadow value of capital.

Caballero and Leahy (1996) clearly regard the combination of perfect competition and constant returns to scale as 'unlikely' to be appropriate assumptions, particularly in the context of large firms. This is an interesting departure from Lucas (1967), who viewed this combination of assumptions as a strength of the underlying q theory. Without linear homogeneity of the net revenue function, the firm has an optimal size, whereas with linear homogeneity of the net revenue function the firm has no optimal size. This reconciles the model with evidence suggesting that changes in firm size are difficult to predict.[21] Recall that these assumptions have been made in the vast majority of structural investment models derived in the absence of fixed costs. Given this, testing a model which maintains these assumptions whilst introducing fixed costs does not seem to be an unreasonable project *a priori*. Whilst linear homogeneity may not hold exactly, it may provide a sufficiently good approximation to be useful, and the fact that this assumption greatly simplifies the characterisation of optimal investment in the presence of fixed adjustment costs would be no small advantage if this is the case. In our view, conclusive evidence that linear homogeneity should be abandoned in the investment literature has not yet been presented.

Abel and Eberly (1996) and Barnett and Sakellaris (1998) have estimated non-linear models relating the investment rate to average q using data on US publicly traded companies. Both studies find significant non-linearities. Barnett and Sakellaris (1998) find a concave relationship, with a flatter slope at higher values of average q. Abel and Eberly (1996) find an S-shaped sigmoidal relationship, with a steeper slope at values of average q around unity and a flatter response at both higher and lower values.[22]

[21] The proposition that firm size follows a random walk is known as Gibrat's Law, and a large literature in empirical industrial organisation has sought to test this hypothesis [see Sutton (1998)]. Whilst the results are somewhat inconclusive, the fact that a large literature exists suggests that the proposition is not without interest, and is not rejected out of hand.

[22] See also Eberly (1997).

At first sight this seems inconsistent with the predictions of the model with non-convex adjustment costs, which implies no response of investment to average q in the intermediate range of inaction, and a response of investment to average q only when average q moves outside this range. Abel and Eberly (1996) suggest that this apparently paradoxical finding can be explained by aggregation over different types of capital goods. The basic idea is that with non-convex adjustment costs, adjustment occurs at both the extensive margin (the decision to invest or not to invest) and the intensive margin (how much to invest or disinvest given that an adjustment is worthwhile). At intermediate values, a rise in average q will trigger adjustment at the extensive margin for some types of capital, with spending on those types of investment increasing from zero to some level that is large enough to justify paying the corresponding fixed cost. At higher values, a further rise in average q produces adjustment only at the intensive margin for all types of capital, so that the response of total investment spending to the change in average q may be flatter.

This argument is ingenious, but it should be noted that there are censoring and measurement error problems that can also account for the non-linear shape found by Abel and Eberly (1996), even if the true relationship is linear. Gross investment in their model is allowed to be negative, and should be so if average q is low enough to justify disinvestment. However the investment data used to estimate this relationship measures only positive capital expenditures, without subtracting any measure of disinvestment. Thus the measure of gross investment used in the empirical work is effectively censored below at zero. This censoring, which is not accounted for in their estimation, can potentially explain a flattening in the response of measured investment to average q at low values of q: once the dependent variable is close to zero, it cannot go much lower, whatever happens to average q. Measurement error in average q can also explain the flattening observed at the upper end of the distribution, particularly if, as seems likely, the measurement error has a higher variance at higher values of average q – for example, if high average q ratios are more likely to be affected by bubbles or fads in share prices. More generally, measurement error in the explanatory variable can change the shape of the underlying relationship, as well as introducing an attenuation bias [see, for example, Chesher (1991)]. As noted earlier, measurement error is likely to be a serious problem with the average q ratio. Bond and Cummins (2001) suggest that noise in share prices can account for findings of non-linearity in this context, when average q is measured using stock market valuations.

The censoring problem noted here may also be more fundamental than it appears. The measure of disinvestment that we would like to subtract from capital expenditures is the replacement cost value of the capital goods that are sold or disposed of. The proceeds from sales of capital goods will underestimate this if there are adverse selection problems in second-hand capital goods markets, and the book value of disposals will underestimate their replacement cost value by an unknown amount depending on the age of the goods that are sold and the inflation rates that have been experienced. Thus whilst it may be important to allow for non-linearities in the relationship between *measured* investment and its determinants, it is not clear that these non-linearities can

identify the shape of the underlying adjustment cost function, at least over the range where disinvestment is important.

Caballero, Engel and Haltiwanger (1995) adopt a somewhat more *ad hoc* approach to test for the importance of fixed adjustment costs.[23] Their empirical analysis is motivated by the following observations. In a model with (possibly asymmetric) fixed costs of adjustment, but no linear or strictly convex components, capital adjustment if it occurs will take the capital stock to the same target level, K_t^*, whether this adjustment occurs from above or from below. Conditional on the fixed cost levels, positive investment will occur when the ratio $Z_t = \frac{K_t^*}{K_t}$ exceeds some critical threshold, and disinvestment will occur when Z_t goes below some threshold. However if the actual fixed cost levels are stochastic and not observed by the econometrician, all we can say is that the probability of observing investment or disinvestment should be increasing in the absolute value of $z_t = \ln Z_t$.

Caballero, Engel and Haltiwanger (1995) implement this test using LRD data on US manufacturing plants. This requires constructing a measure of the target capital stock (K_t^*), and hence the 'gap' (z_t). Although in general this target will not be equal to the optimum capital stock that the firm would choose in the absence of adjustment costs, Caballero, Engel and Haltiwanger (1995) base their measure of the target on the predicted values from a static demand for capital equation of the form (2.13). They find that the probability of observing investment is clearly increasing with positive values of z_t, consistent with the presence of the fixed costs. They also find an asymmetry, with the probability of observing disinvestment remaining low even at large negative values of z_t, although as they note this asymmetry could reflect mis-measurement of disinvestment in the data. This 'gap' methodology has also been criticised by Cooper and Willis (2004), who suggest that the findings may be sensitive to mis-specification of the target level to which the actual capital stock is assumed to adjust.

Finally in this section we should mention interesting recent work by Cooper and Haltiwanger (2006). They also use US plant level data to investigate capital stock adjustment, but rather than directly estimating an investment equation they use indirect inference to estimate parameters of a general adjustment cost function. In essence, this asks what forms of adjustment costs are required to match the non-linear relationship between investment and profitability found in their data set. Perhaps not surprisingly, they find that a general specification of adjustment costs is required to fit this relationship, combining both convex and non-convex components with irreversibility. This finding, like the descriptive evidence on plant-level investment discussed earlier, casts doubt on the structural econometric investment models that have maintained much simpler adjustment cost specifications. The challenge remains to develop structural investment equations that are consistent with richer forms of adjustment costs, without requiring the net revenue function to be homogeneous of degree one.

[23] See also Cooper, Haltiwanger and Power (1999) and Anti Nilsen and Schiantarelli (2003).

3.6. Reduced form models

The actual adjustment process is likely to be extremely complex, particularly when we consider that the data on investment at the firm level are an aggregate over many types of capital goods, and possibly over multiple plants. The structural models of investment dynamics that have been proposed to date have not been conspicuously successful in characterising this adjustment process, possibly because they have neglected these aggregation issues, and/or the implications of some of the non-convexities discussed in the previous section. Nevertheless it is clear that capital cannot be adjusted costlessly and immediately, so it is not appropriate to resort to static models.[24] An intermediate possibility is to rely on dynamic econometric specifications that are not explicitly derived as optimal adjustment behaviour for some particular structure of adjustment costs. A favourable interpretation of such reduced form models is that they represent an empirical approximation to some complex underlying process that has generated the data. A less favourable interpretation is that they compound parameters of the adjustment process with parameters of the expectations-formation process, and are subject to the Lucas (1976) critique. In any case it is useful to be aware of the form of these models, and some of their limitations.

One approach which has been widely used in the investment literature is based on first-differencing a static factor demand model to obtain an investment equation, for example

$$\left(\frac{I_t}{K_{t-1}}\right) - \delta \approx \Delta k_t = \Delta k_t^* \tag{3.20}$$

where k_t^* is the logarithm of the optimal capital stock, which may for example be given by Equation (2.13). If k_t^* is log-linear in the level of output, this leads to versions of the popular accelerator model, in which investment is related to output growth. Recognising that the actual capital stock does not adjust fully and immediately to changes in the desired level, so-called flexible accelerator models introduce distributed lags in Δk_{t-s}^* and possibly Δk_{t-s}. This gives a dynamic specification of the form

$$a(L)\Delta k_t = b(L)\Delta k_t^*, \tag{3.21}$$

where $a(L)$ and $b(L)$ are polynomials in the lag operator (i.e. $L^s x_t = x_{t-s}$). Flexible accelerator models of this type have been estimated using firm data by, for example, Eisner (1977) and Mairesse and Dormont (1985).

An alternative is to specify a simple partial adjustment model for the level of the capital stock, such as

$$\left(\frac{I_t}{K_{t-1}}\right) - \delta \approx \Delta k_t = \theta\left(k_t^* - k_{t-1}\right), \tag{3.22}$$

[24] Unless perhaps one is willing to treat all variables of interest as co-integrated non-stationary variables and one has sufficiently long time series to appeal to co-integration results [cf. Engle and Granger (1987)].

in which some constant fraction θ of the gap between the actual and desired levels of the capital stock is closed in each period. This is obviously very restrictive, but a more flexible dynamic adjustment model which nests both partial adjustment and accelerator models as special cases is an error correction model, such as

$$\alpha(L)\Delta k_t = \beta(L)\Delta k_t^* + \theta\left(k_{t-s}^* - k_{t-s-1}\right) \tag{3.23}$$

where $\alpha(L)$ and $\beta(L)$ are again polynomials in the lag operator, the form of which can be chosen empirically.[25] Error correction models were introduced into the investment literature by Bean (1981), and have been considered in the context of firm data by Bond, Harhoff and Van Reenen (2003) and Bond et al. (2003).

The connection between error correction models and co-integration techniques have popularised these adjustment models in the time series literature, but the error correction model was introduced into econometrics long before the literature on co-integration developed.[26] In fact, the error correction model is nothing more than a particular parameterisation of an autoregressive distributed lag (ADL) model. For example, the ADL(1, 1) model

$$k_t = \alpha_1 k_{t-1} + \beta_0 k_t^* + \beta_1 k_{t-1}^* \tag{3.24}$$

can always be re-parameterised as

$$\Delta k_t = -\beta_1 \Delta k_t^* + (1-\alpha_1)\left(k_t^* - k_{t-1}\right) \tag{3.25}$$

under the long-run proportionality restriction $(\frac{\beta_0+\beta_1}{1-\alpha_1}) = 1$.[27]

Implementation of the error correction model will require a specification for the target level of the capital stock. Subject again to the observation that the target capital stock in the presence of adjustment costs is not necessarily equal to the desired capital stock in the absence of adjustment costs, this can be based on a static factor demand specification such as (2.13). For example, combining (2.13) and (3.25) gives

$$\begin{aligned}\Delta \ln K_t = {} & (1-\alpha_1)\sigma \ln a_K \left(1 - \frac{1}{\eta^D}\right) - \beta_1 \Delta \ln Y_t + \beta_1 \sigma \Delta \ln\left(\frac{r}{p}\right)_t \\ & + (1-\alpha_1)(\ln Y_t - \ln K_{t-1}) - (1-\alpha_1)\sigma \ln\left(\frac{r}{p}\right)_t .\end{aligned} \tag{3.26}$$

Notice that these models can be used to estimate the long-run elasticity of the capital stock with respect to the user cost of capital, whilst allowing for the fact that this adjustment does not occur immediately. In principle these models can be extended to

[25] See Hendry (1995, Chapter 7) for further discussion of alternative dynamic models.

[26] See, for example, Sargan (1964) and Davidson et al. (1978).

[27] This restriction can be tested by including an additional term in either k_t^* or k_{t-1} on the right-hand side of (3.25).

incorporate non-linear and asymmetric dynamics, for example by allowing the parameters α_1 and β_1 to take different values depending on the sign of $(k_t^* - k_{t-1})$ and whether the absolute value of this gap is large or small.

Since these reduced form models are empirical generalisations of static factor demand specifications, and the adjustment cost models considered previously are theoretical generalisations of static factor demand theory, it is no surprise that the two approaches are related. Nickell (1978, Chapter 11) shows that for a symmetric, quadratic adjustment cost function, the level of investment can be obtained approximately as

$$I_t - \delta K_{t-1} = \Delta K_t \approx \phi(\widehat{K}_t - K_{t-1}), \tag{3.27}$$

where

$$\widehat{K}_t = (1-\gamma)E_t\left[\sum_{s=0}^{\infty} \gamma^s K_{t+s}^*\right] \tag{3.28}$$

and K_t^* is the capital stock level that the firm would have chosen in the absence of adjustment costs. Thus the investment decision is approximately described by a partial adjustment mechanism, in which the 'target' level of the capital stock is itself a function of both current and expected future levels of the static optimum.[28] Nickell (1985) characterises restrictive conditions for the K_t^* process under which this generates either a partial adjustment or an error correction model relating investment to $(K_t^* - K_{t-1})$. Whilst this relationship is of interest, it does not provide a very appealing motivation for the use of these reduced form empirical models. If this structure for the adjustment costs is taken seriously, we have seen how this structure can be identified and tested more directly. The attraction of the reduced form models is that they may provide an empirical approximation to a much more complex adjustment process, whose structure we have not yet been able to characterise satisfactorily as the outcome of a richer dynamic optimisation problem.

The potential disadvantages of the reduced form approach can now be illustrated as follows. Suppose for simplicity that the structural adjustment process was characterised by

$$I_t = \alpha K_{t-1} + \beta_0 Y_t + \beta_1 E_t[Y_{t+1}], \tag{3.29}$$

which is a special case of (3.27) in which we have omitted expected values beyond period $t+1$, and assumed that the static optimum capital stock is proportional to output. Suppose also that expectations of future output are formed according to

$$E_t[Y_{t+1}] = \pi_0 Y_t + \pi_1 Y_{t-1} + \pi_2 X_t, \tag{3.30}$$

where X_t is a vector containing any additional variables that happen to be useful in forecasting future output. Then substituting (3.30) into (3.29) we obtain the reduced

[28] Recall that this dependence of current investment decisions on expected future profitability was also evident from the first-order conditions (see Equation (3.3)).

form investment equation

$$I_t = \alpha K_{t-1} + (\beta_0 + \beta_1\pi_0)Y_t + \beta_1\pi_1 Y_{t-1} + \beta_1\pi_2 X_t. \tag{3.31}$$

This shows how the reduced form models compound the parameters of the structural adjustment process (α, β_0, β_1) with the parameters of the expectations-formation process (π_0, π_1, π_2).[29] This has two potentially important consequences. First, if the parameters of the expectations-formation process are not stable, then the parameters of the reduced form investment equation will also be unstable, even though the parameters of the structural adjustment process may have been constant. For example, if there is a structural break in the process generating output, perhaps as a result of entry into the market or some change in macroeconomic policy, this will induce parameter instability in the reduced form investment model. This simply illustrates that the reduced form models are subject to the Lucas (1976) critique. In principle this is not the case for the structural investment models we considered in Sections 3.1–3.3, whose parameters are 'deep' parameters describing the adjustment cost technology (see (3.10) or (3.13)), and which are expected to be invariant to changes in the processes generating, for example, output and the user cost. In practice, this claim relies on these structural models being correctly specified, and even then we may find that the parameters of the adjustment cost function are not constant over time.

The second consequence is that X_t variables will appear to be significant in reduced form equations like (3.31), even though they play no role in the structural model for investment, and their only role in the reduced form equation is to help to forecast future values of the fundamental determinants of investment. This is clearly problematic if we want to draw any inferences about the nature of the underlying structural model. For example, finding that financial variables have significant coefficients in a reduced form investment equation does not identify whether these financial variables are important structural determinants of investment spending – perhaps as a result of financing constraints – or whether they simply help to forecast future output or profitability. We discuss some possible solutions to this identification problem further in Section 6.2 below. Here we note that whilst the problem is particularly transparent in the context of reduced form models, a similar issue will affect structural models that are not correctly specified, and which therefore do not fully control for all influences of expected future output or profitability on the level of current investment.

[29] If the structural adjustment process were really as simple as Equation (3.29), we could of course substitute the realised value of future output for the one-step ahead expectation. However this will not generally be possible when the structural model relates current investment to expectations of output or profitability in the distant future, as suggested by (3.27) or (3.3).

4. Econometric issues

4.1. Stochastic specification and identification

One important issue in the implementation of econometric investment and employment models concerns the sources of stochastic error terms. In this section we discuss in particular the stochastic specifications that have been considered in microeconometric applications of the Q and Euler equation models of investment, and their implications for the consistent estimation of these models.

The Q model derived in Equation (3.10) is a deterministic relationship between the investment rate and the Q variable. The intercept in this equation is a parameter of the adjustment cost function, interpreted as a 'normal' rate of investment at which costs of adjustment are zero. The standard way of introducing stochastic variation into the Q model is to treat this parameter as stochastic, and to interpret the error term in the Q investment equation as reflecting shocks to the adjustment cost function. Simply replacing the constant a with $a_{it} = a + e_{it}$ for firm i in time period t gives the econometric model

$$\left(\frac{I}{K}\right)_{it} = a + \frac{1}{b} Q_{it} + e_{it}, \tag{4.1}$$

where e_{it} is an additive shock to the 'normal' rate of investment, or equivalently an additive shock to marginal adjustment costs.

This approach is convenient, if somewhat *ad hoc*. Nevertheless it has several implications which should be taken seriously when estimating the Q model. Perhaps most importantly, it implies that Q_{it} should be an endogenous variable in the econometric model (4.1). Current shocks to adjustment costs will affect the current period's net revenue (Π_{it}), and therefore the current value of the firm (V_{it}).[30] This endogeneity of Q_{it} will therefore need to be taken into account in order to obtain consistent estimates of the adjustment cost parameters.

A second implication is that technological shocks need not be statistical innovations. There may be permanent differences across firms in their 'normal' investment rates, and there may be common trends in the nature of adjustment costs that affect all firms in the same way, perhaps as a result of business cycle fluctuations. Thus for estimation of the Q model using company panel data it is not inconsistent with the underlying theory to include firm-specific and time-specific error components. Simply letting $e_{it} = \eta_i + \zeta_t + \varepsilon_{it}$ gives the error components specification

$$\left(\frac{I}{K}\right)_{it} = a + \frac{1}{b} Q_{it} + \eta_i + \zeta_t + \varepsilon_{it}. \tag{4.2}$$

[30] See Equation (3.8). Alternatively, the current adjustment cost shock affects the current marginal revenue product of capital, and therefore the current shadow value of capital (λ_{it}). See Equation (3.3).

Moreover, there is no compelling reason for the idiosyncratic, time-varying component of adjustment cost shocks (ε_{it}) to be serially uncorrelated. For example, if these shocks follow an $AR(1)$ process, $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + \nu_{it}$, with ν_{it} serially uncorrelated, then we obtain the dynamic specification

$$\left(\frac{I}{K}\right)_{it} = a(1-\rho) + \rho\left(\frac{I}{K}\right)_{i,t-1} + \frac{1}{b}Q_{it} - \frac{\rho}{b}Q_{i,t-1} + \eta_i(1-\rho) + \zeta_t - \rho\zeta_{t-1} + \nu_{it}, \tag{4.3}$$

i.e. a dynamic model relating the current investment rate to both current and lagged Q and the lagged dependent variable, subject to a non-linear common factor restriction. More generally, the underlying theory does not rule out a dynamic relationship between investment rates and the Q variable, provided these dynamics are consistent with some serial correlation process in the adjustment cost shocks.

A second potential source of stochastic variation in the Q model is measurement error. In view of both the assumptions required to measure marginal q, using either the average q ratio or an auxiliary econometric forecasting model, and the limitations of most publicly available datasets,[31] the likelihood of significant measurement error in the Q variable does indeed seem overwhelming. Denoting the true value of the explanatory variable on the right-hand side of (3.5) or (3.10) by Q^*_{it}, and the measured value by $Q_{it} = Q^*_{it} + m_{it}$, where m_{it} is an additive measurement error, then gives an econometric model of the same form as (4.1), with $e_{it} = -\frac{m_{it}}{b}$. Again a principal implication is that measured Q_{it} will be correlated with the error term, and this endogeneity should be allowed for in estimation.[32] Moreover, the measurement error may also have firm-specific, time-specific and serially correlated components. Unfortunately the residual measurement error component may still not have properties that are convenient for estimation, and alternative techniques such as those considered in Erickson and Whited (2000), or alternative measures such as those considered in Bond and Cummins (2001), may still be required to identify the parameters of the underlying model.

In contrast to the Q model, the Euler equation approach considered in Section 3.3 has an intrinsically stochastic specification when the one-step ahead expected values are replaced by their realised values. Assuming the real discount factor ψ_{t+1} is a constant parameter ψ and denoting $(\frac{I}{K})_{i,t+1} = E_{it}[(\frac{I}{K})_{i,t+1}] + \epsilon_{i,t+1}$, where $\epsilon_{i,t+1}$ is the error made by firm i when forecasting its period $t+1$ investment rate using information available in period t, the Euler equation in (3.13) becomes

$$\left(\frac{I}{K}\right)_{it} = a(1-\psi) + \psi\left(\frac{I}{K}\right)_{i,t+1}$$

[31] See Section 5 below.

[32] Though it is perhaps worth noting that whilst the theory predicts a positive correlation between current Q_{it} and adjustment cost shocks in the absence of measurement error, the correlation between measured Q_{it} and $-\frac{m_{it}}{b}$ will be negative. Thus the direction of the potential simultaneity bias becomes ambiguous if both adjustment cost shocks and measurement errors are present.

$$+\frac{1}{b}\left[\left(\frac{\partial F}{\partial K_{it}}\right)-\left(\frac{\partial G}{\partial K_{it}}\right)-\left(\frac{r}{p}\right)_{it}\right]-\psi\epsilon_{i,t+1}. \tag{4.4}$$

The forecast error $\epsilon_{i,t+1}$ will certainly be correlated with $(\frac{I}{K})_{i,t+1}$, but under weak rational expectations should be orthogonal to information available in period t. One implication is that this forecast error cannot be serially correlated, and cannot contain a permanent firm-specific component.[33] On the other hand, since firms are likely to be subject to common shocks reflecting, for example, business cycle surprises, these forecast errors are likely to be correlated across firms in each period. This suggests that identification of the Euler equation may be problematic without long time series of data for individual firms, since independence of the error terms across individuals is typically required for consistent estimation in panels when the number of time periods is fixed. The inclusion of time dummies will only permit consistent estimation in the special case where the effects of aggregate shocks are perfectly correlated across individual firms. Of course, the assumption that technological shocks and measurement errors are independent across firms (conditional on time dummies) may also be problematic in the context of the Q model and other factor demand specifications, but this assumption is particularly difficult to reconcile with the underlying economic structure when the stochastic disturbances contain a forecast error.[34]

It should also be noted that unobserved heterogeneity in the real discount factor ψ could undermine the identification of the Euler equation. Measured heterogeneity in ψ, reflecting for example differences in the mix of capital assets used by different firms or differences in their required rates of return, can be allowed for by including suitable interaction terms in the estimated model. Unobserved heterogeneity in ψ is problematic because, as is clear from (4.4), this interacts with the endogenous variable $(\frac{I}{K})_{i,t+1}$. Unrestricted unobserved heterogeneity (ψ_{it}) would therefore leave the Euler equation unidentified. Restricting the unobserved heterogeneity to be permanent and firm-specific $(\psi_{it}=\psi_i)$ would allow consistent estimation if long time series data were available, but even this restriction would leave an identification problem in panels with a small number of time periods. As has been emphasised by Pesaran and Smith (1995), heterogeneous slope coefficients will invalidate the instruments typically used to estimate dynamic models from short panels.[35] Again we should acknowledge that a similar

[33] Although firm-specific measurement errors or omitted variables *may* rationalise firm-specific effects in empirical applications of Euler equation models of factor demand.

[34] Interested readers are referred to Altug and Miller (1990, 1991) and Blundell, Bond and Meghir (1996) for further discussion of this issue.

[35] Although it is less clear from (4.4), a similar issue can arise if there is unobserved heterogeneity in the adjustment cost parameter a. The source of this potential problem is the $[(\frac{\partial F}{\partial K_{it}})-(\frac{\partial G}{\partial K_{it}})]$ term. If, as in Bond and Meghir (1994), linear homogeneity of the production and adjustment cost functions is used to measure this marginal product, the effect is to introduce a linear term in $a(\frac{I}{K})_{it}$. Indeed it should be noted that this form of the Euler equation is inconsistent with the presence of the kind of adjustment cost shocks that are typically used to motivate stochastics in the Q model.

issue may arise in the context of the Q model if the adjustment cost parameter (b) is itself heterogeneous across firms.

The potential sources of stochastic variation in reduced form investment equations are less transparent. If these models are regarded as empirical approximations to some complex adjustment process, then the residual can be viewed as an approximation error, but this gives little guidance as to its statistical properties. One approach in this context is to regard certain properties of the error term (e.g. lack of serial correlation) and certain properties of the model parameters (e.g. stability) as features of a desirable approximation to the data generation process, and to treat these properties as key objectives of an empirical specification search. Since this raises issues that are common to reduced form models in a wide range of econometric applications, we refer interested readers to Hendry (1995) for further discussion.

4.2. Estimation

Many of the issues that arise in estimating dynamic factor demand equations using company panel data can be illustrated by considering a model of the form

$$y_{it} = \alpha y_{i,t-1} + \beta x_{it} + u_{it} \quad \text{for } i = 1, \ldots, N \text{ and } t = 2, \ldots, T,$$
$$u_{it} = \eta_i + v_{it} \tag{4.5}$$

where i indexes firms and t indexes time, and η_i is an unobserved firm-specific effect.[36] For example, if y_{it} is the investment rate, x_{it} is Q_{it} as defined in (3.10) and $\alpha = 0$ then we have the basic Q model, whilst if $\alpha \neq 0$ and x_{it} is a vector containing both Q_{it} and $Q_{i,t-1}$ we have the Q model with an $AR(1)$ component to the adjustment cost shock. The Euler equation model (4.4) has this form if we normalise on $(\frac{I}{K})_{i,t+1}$ rather than $(\frac{I}{K})_{it}$,[37] and the reduced form models considered in Section 3.6 are typically just more general dynamic equations of this type. For simplicity, this section will focus on the case where x_{it} is a scalar.

Since the available company panels typically contain a large number of firms observed for a relatively small number of time periods, we will concentrate on the estimation issues that arise when N is large and T is small, assuming that the error term v_{it} is distributed independently across firms.[38] OLS will give biased parameter estimates since $y_{i,t-1}$ is necessarily correlated with η_i (and x_{it} may be), and the Within estimator will give biased parameter estimates for small T since $\tilde{y}_{i,t-1}$ is necessarily correlated with $\tilde{u}_{it}$ (and $\tilde{x}_{it}$ may be).[39,40]

[36] See Arellano and Honoré (2001) for a more detailed discussion of the methods we review here.

[37] Similar estimation issues will arise whichever normalisation is adopted.

[38] Unless otherwise indicated, asymptotic properties will hold as N goes to infinity for fixed T.

[39] The Within estimator is obtained as OLS after subtracting firm-means of each variable, so that $\tilde{y}_{it} = y_{it} - \frac{1}{T-1}\sum_{s=2}^{T} y_{is}$ and $\tilde{y}_{i,t-1} = y_{i,t-1} - \frac{1}{T-1}\sum_{s=1}^{T-1} y_{is}$. This is equivalent to the Least Squares Dummy Variables (LSDV) estimator, obtained by including a dummy variable for each firm.

[40] In the special case where $\beta = 0$, we can say that the OLS estimate of α will be biased upwards [Hsiao (1986)] and the Within estimate of α will be biased downwards [Nickell (1981)]. Informally these estimates

If x_{it} is strictly exogenous with respect to u_{is}, in the sense that $E[x_{it}u_{is}] = 0$ for all s, t (or some strictly exogenous instrument z_{it} is available), the parameters (α, β) can be estimated consistently by using the vector $(x_{i1}, \ldots, x_{iT})$ – or $(z_{i1}, \ldots, z_{iT})$ – as instrumental variables for each of the levels equations in (4.5). If x_{it} is correlated with η_i but strictly exogenous with respect to v_{is} (or some instrument z_{it} with these properties is available), the parameters (α, β) can be estimated consistently by taking first-differences of (4.5) and then using the vector $(x_{i1}, \ldots, x_{iT})$ – or $(z_{i1}, \ldots, z_{iT})$ – as instrumental variables for each of the resulting first-differenced equations. More commonly, in the absence of strictly exogenous instruments, identification of (α, β) relies on assuming some limited serial correlation in the v_{it} disturbances.

For example, consider the case in which x_{it} is correlated with both η_i and v_{it}. Assuming that $E[y_{i1}v_{it}] = E[x_{i1}v_{it}] = 0$ for $t = 2, \ldots, T$ and that $E[v_{is}v_{it}] = 0$ for $s \neq t$ yields the moment conditions

$$E[w_{i,t-s}\Delta v_{it}] = 0 \quad \text{for } s \geqslant 2 \text{ and } t = 3, \ldots, T, \tag{4.6}$$

where $w_{it} = (y_{it}, x_{it})$. This allows the use of lagged values of endogenous variables dated $t - 2$ and earlier as instrumental variables for the equations in first-differences.[41] For $T > 3$ the parameters are over-identified, and alternative Generalised Method of Moments (GMM) estimators are defined by different ways of weighting the moment conditions. If instead of assuming no serial correlation we allow v_{it} to be $MA(q)$, the implication is that only lagged endogenous variables dated $t - 2 - q$ and earlier can be used as instruments.

This first-differenced GMM estimator will provide consistent estimates of the parameters (α, β) as the number of firms becomes large. In some contexts these estimators have good finite sample properties, but this will not be the case when the lagged values of the series are only weakly correlated with subsequent first-differences. When the instruments available are weak, the GMM estimator can exhibit large finite sample biases, as well as imprecision.[42] In the context of Equation (4.5) this is particularly likely to be a problem when the individual series for y_{it} and x_{it} are highly persistent, and when the time series dimension of the panel is very short.[43]

Alternative estimators are available that have better small sample properties in these cases, although they have been less commonly used in applied work to date. Alonso-Borrego and Arellano (1999) propose a symmetrically-normalised GMM estimator that

provide guidance about the likely range of values for α, so that candidate consistent estimators which lie well outside this range can often be regarded with suspicion.

[41] See Holtz-Eakin, Newey and Rosen (1988) and Arellano and Bond (1991). $x_{i,t-1}$ could be used as an additional instrument if x_{it} is predetermined with respect to v_{it}. Additional instruments are available for the levels equations if x_{it} (or Δx_{it}) is uncorrelated with η_i.

[42] For general results, see Nelson and Startz (1990a, 1990b) and Staiger and Stock (1997). For the case of dynamic panel data models, see Blundell and Bond (1998).

[43] In the special case with $\beta = 0$, Blundell and Bond (1998) show that the first-differenced GMM estimator of α has a serious downward bias in these cases.

is median-unbiased even in the case of weak instruments. Ahn and Schmidt (1995) note that estimators based on the linear moment conditions (4.6) are not efficient under the standard 'error components' assumption that $E(\eta_i v_{it}) = 0$, and that additional non-linear moment conditions are available in this case. The resulting non-linear GMM estimator is asymptotically efficient relative to the linear first-differenced GMM estimator, and can be expected to have better small sample properties.[44]

Estimators with better properties can also be obtained in some contexts if one is able to impose more restrictive assumptions than those considered above. Arellano and Bover (1995) note that Δx_{it} may be uncorrelated with the unobserved firm-specific effects, even when the level of x_{it} is correlated with η_i. Combined with the assumption of limited serial correlation, this allows suitably lagged values of the first-differences $\Delta x_{i,t-s}$ to be used as additional instruments for the equations in levels. These additional moment conditions are over-identifying restrictions that can be tested, and where they are valid they can significantly improve on both the asymptotic and small sample properties of the first-differenced GMM estimator, particularly when the parameters (α, β) are only weakly identified from the first-differenced equations. Blundell and Bond (1998, 2000) note that if the initial conditions y_{i1} also satisfy the stationarity restriction $E[\Delta y_{i2}\eta_i] = 0$ then suitably lagged values of $\Delta y_{i,t-s}$ as well as $\Delta x_{i,t-s}$ are available as instruments for the levels equations.[45] The resulting linear estimator, which combines equations in levels with equations in first-differences and which they label 'system GMM', is shown to provide dramatic gains both in asymptotic efficiency and in small sample properties, compared to both the linear first-differenced GMM estimator and to the non-linear GMM estimator of Ahn and Schmidt (1995).

Another potentially important issue in the context of panel data on companies or plants is that of non-random entry and exit. Entry into a panel of establishments may reflect the decision by a firm to enter a particular market, whilst entry into a panel of firms may reflect other economic choices such as the decision to obtain a stock market listing. Exit from a panel of plants may reflect plant closure or acquisition, whilst exit from a panel of companies may reflect mergers, takeovers and bankruptcies. Many of these economic events are potentially correlated with shocks that affect investment and employment decisions.

Non-random entry into the sample does not present a serious estimation problem, however, since the entry decision is a function of variables dated at the time of entry, which can be regarded as fixed over the subsequent sample period. The variables

[44] Again for the special case with $\beta = 0$, Ahn and Schmidt (1995) show that the gain in asymptotic efficiency is largest in the cases where α is only weakly identified from the linear moment conditions (4.6) – i.e. when α is high and T is small.

[45] In the special case with $\beta = 0$, this initial condition restriction requires that the series $(y_{i1}, \ldots, y_{iT})$ have a constant first moment for each firm. In the multivariate context, Blundell and Bond (2000) show that a sufficient condition is for the series (y_{it}, x_{it}) to both have stationary means. However, this is not necessary. For example, both y_{it} and x_{it} may have non-stationary means provided that the Δx_{it} series is always uncorrelated with η_i and the conditional model (4.5) has generated the y_{it} series for a sufficiently long pre-sample period.

which determine entry may be correlated with the firm-specific effects in factor demand equations, but this fixed correlation merely shifts the firm-specific effects and can be controlled for using the estimation methods described above. The determinants of entry may also be correlated with subsequent shocks to the factor demand equation if the latter are serially correlated, but this correlation can be controlled for by dropping a limited number of initial periods.

Non-random exit from the sample could cause a potentially more serious attrition bias, that would only be corrected by controlling for firm-specific effects under very restrictive assumptions. This would remain an issue whether estimation is based on the 'unbalanced' panel, including those firms which exit, or the 'balanced' panel containing only the subset of firms which survive through to the end of the sample period. The development of tests and controls for attrition bias in panels of firms and plants is an important area for future research, although it should be recognised that this may not be a purely statistical issue. For example, in principal-agent models of company behaviour with incomplete contracts, where managers retain some flexibility not to pursue shareholder value maximization, it may be the case that investment decisions are directly influenced by the risk of being taken over or going bankrupt, since managers who dislike these events may choose actions to reduce their risk. Thus the nature of the exit processes may affect the specification of factor demand models, and not only the estimation methods required for consistent estimation.

5. Data

The advantages of microeconomic data sets are that data is often available for a large number of individual firms or plants, and aggregation problems are generally reduced compared to industry level or aggregate data sets. Moreover, it is possible to move beyond a representative firm framework, and to test models that imply heterogeneous behaviour across firms, as for example may occur in the presence of financing constraints. The disadvantages are that the available measures of factor inputs and outputs are often crude, and key variables like factor prices are generally not measured at the firm or plant level. This is a major reason why much of the microeconometric literature on investment and employment has focused on issues such as the nature of adjustment dynamics and the presence of financing constraints, which can be investigated without microeconomic variation in factor prices.

Until very recently, most microeconometric factor demand studies relied on publicly available company data sets, generally obtained from company accounts. Examples include Compustat data for US firms and Datastream data for UK firms. Company accounts are not produced for the benefit of econometric research, and the measures that are directly available are often inappropriate for testing economic models. Measures based on recorded cash flows are likely to be most reliable, but even these can present problems for econometric research. For example, whilst data on sales may be accurate,

there may be too little information on changes in inventories to infer the value of production. Similarly there may be insufficient information on the cost of current inputs to infer value-added. However flow measures that are based on changes in accounting valuations of assets and liabilities present more severe measurement issues, as does the use of the book values of these stocks themselves.

These book values of assets and liabilities are generally based on historic cost valuations, which may deviate substantially from current economic values in the case of long-lived capital assets and long-term debts. The historic cost valuation of the firm's capital stock is based on the prices at which assets were originally purchased, and so neglects both general price inflation and relative price changes over the intervening period. A related concern affects the depreciation deductions reported in company accounts. Even in commercial accounts, these may be based on cautious assumptions about the length of useful asset lives, and in some cases the only information available is based on depreciation rules required for tax purposes. For these reasons economic researchers have often preferred to construct their own capital stock estimates, based on cumulating the observed investment flows in a perpetual inventory formula that can allow for inflation and alternative estimates of economic depreciation rates. A similar problem affects the valuation of inventories, although here it is important to know the valuation method that has been used to construct the company accounts, which varies across countries and may vary across firms within the same country. The historic cost valuation of debt is based on the amounts borrowed and not yet repaid. The market value of the firm's outstanding debt may be different if interest rates have changed over the intervening period. Again economic researchers have sometimes preferred to construct alternative estimates of the market value of debt, although this is problematic without knowledge of the maturity structure. Many of these valuation problems come together in accounting measures of profit, with some of the principal concerns being the deduction of nominal rather than real interest payments, the inclusion of inflationary gains on holdings of inventories, and the use of historic cost depreciation charges.[46] Not to mention the possibility that the timing of some charges against profits may be manipulated to manage the release of news about the company to the financial markets.

Of particular concern in the factor demand context is the poor quality of accounting information on the use of various inputs. For example, company accounts may report expenditures on direct purchases of fixed capital, but may give little information on the breakdown of these expenditures by type of asset, or may provide little information on the value of fixed capital obtained through the acquisition of other firms, or may provide little information on the value of fixed capital sold or scrapped. Thus the available data on investment may be subject to measurement errors whose importance may differ across sectors and over the business cycle. As we discussed in Section 3.5, the absence of good data on disposals of capital may be particularly important when testing for asymmetries in upward and downward adjustments. The available data on employment in company accounts is often limited to a snapshot of the number of employees:

[46] See Edwards, Kay and Mayer (1987) for a comprehensive discussion of accounting measures of profit.

it is comparatively rare to have information on hours worked, skill composition or flow measures of hiring and separations, all of which would be desirable when investigating the structure of labour demand. In company accounts data it is also unusual to have data on the usage of factors other than capital and labour.

A distinct source of micro data is provided by the large establishment level data sets that are compiled by government statistical agencies in order to estimate the aggregate levels of production, investment and employment in different sectors of the economy. Typically these are populations of all establishments above a certain size threshold, and stratified random samples of smaller establishments. An establishment may comprise one or more plants, which may account for all or part of a firm's activities. In the last ten years, these data sources have increasingly been made available for empirical research. Examples include the Longitudinal Research Database (LRD) in the US; the Annual Respondents Database (ARD) in the UK; the Norwegian data used by Anti Nilsen and Schiantarelli (2003); and the French data underlying the Abowd, Kramarz and Margolis (1999) study.[47]

There are several advantages of these establishment data sets over company accounts data. First, they provide data at a more disaggregated level, with the unit of observation generally being a plant or plants in the same geographical location. The LRD also provides quarterly observations. Secondly, these data sources provide far more coverage of the activities of smaller firms than do most company databases, which are often limited to publicly traded firms. Thirdly, the establishment data usually disaggregate information on factor inputs to a greater degree, allowing some consideration of multiple types of capital and labour. On the other hand, some variables may only be measured at the level of the firm, such as stock market valuations and tax payments. Coverage may also be limited to manufacturing or production industries, excluding the service sector which accounts for a large and growing share of investment and employment in developed economies.

As we have stressed in this chapter, a major problem facing microeconometric research on factor demand is the absence of comprehensive micro data on factor prices. This is partly a conceptual problem and not just a data limitation: if each factor of production were bought and sold in a single competitive market, then all firms would face the same price for each factor, and there would be no cross-section variation in factor prices that could be used to identify factor demand parameters. In practice there may be regional and sectoral differences in factor prices, although this is less useful in the context of large firms which may operate in several locations and industries. Thus considerable ingenuity is typically required to identify compelling sources of exogenous price variation. Examples from the investment literature include variation in the effects of taxes across different firms, either because they use different mixes of capital subject to different tax treatments, as in Auerbach and Hassett (1992), Cummins, Hassett and Hubbard (1994) and Chirinko, Fazzari and Meyer (1999); or because they

[47] See Abowd and Kramarz (1999) for more details.

are affected differently by non-linearities in the tax system, as in Devereux (1989) and Devereux, Keen and Schiantarelli (1994). Examples from the labour demand literature include regional variation in minimum wage legislation, and variation in the extent of unionisation.

Our final remarks on sources of microeconomic data for factor demand studies concern different levels of aggregation at which data is available. Large companies may own many subsidiary firms, so that even with company accounts data there is an important distinction between the *unconsolidated* accounts reported by individual subsidiaries, and the *consolidated* accounts reported by the group as a whole. Individual firms may also operate multiple plants, either in different locations or performing different activities. Data at different levels of aggregation may be most suitable for addressing different questions. For example, if the objective is to uncover the structure of the adjustment cost technology, then it seems appropriate to use the most disaggregated data available, to avoid the tendency for discrete adjustments to be smoothed by aggregation across plants or types of capital or labour.[48] As noted previously, commonly used micro data on companies or large plants may still be too aggregated for this purpose. However if the objective is to investigate the impact of financing constraints, it may well be appropriate to consider consolidated data for the company as a whole: even in the presence of a financing constraint, the spending by an individual subsidiary firm or plant may not be constrained by its own cash flow, since the company can reallocate financial resources between different parts of the group. More generally, the relevant locus for at least some aspects of corporate decision making may be at the level of the firm rather than the plant, and it is worth noting that value-maximising behaviour at the level of the company need not imply value maximisation for individual subsidiaries or plants. There are also likely to be important advantages from combining both establishment and firm level data, for example to investigate the impact of aggregation on investment and employment dynamics. Finally, as we noted in the introduction, important aspects of aggregate adjustment may take place through the entry and exit of individual firms or plants, so evidence from microeconomic data will not necessarily provide the correct answers to macroeconomic questions.

6. Topics in investment

6.1. Some basic findings

Many studies have used company panel data to evaluate the Q model of investment. This model appeared particularly well suited to company data sets, since stock market valuations are readily and accurately measured for companies with publicly traded shares, and in contrast to the user cost of capital, there is rich cross-section variation in the average q ratio that should help to identify the model. There was also initial optimism that

[48] See, for example, Hamermesh (1989) and Anti Nilsen and Schiantarelli (2003).

some of the apparent failings of the Q model that had been reported in earlier studies using aggregate data may be due to aggregation problems that arise when investment rates and average q, both specified in ratios, are constructed using aggregate data.

In fact, as we noted in Section 3.1, most of the empirical problems found with the aggregate data have been reproduced in microeconometric studies. These include very low coefficients on the Q variable, suggesting incredibly high marginal costs of adjustment, and violation of the prediction that Q should be a sufficient statistic for investment. In most micro studies, additional variables such as cash flow or sales have been found to be informative after controlling for Q, and in some cases the Q variable becomes insignificant when these other variables are added to the empirical model. Similar findings have been reported independently using data for a wide variety of countries and time periods, and this has also been the case in the relatively small number of studies that have recognised the potential importance of endogeneity and (transient) measurement error in average q. For example, both Hayashi and Inoue (1991) and Blundell et al. (1992) have used first-differenced GMM estimators of the type described in Section 4.2 to estimate versions of the Q model, for panels of Japanese and UK listed manufacturing companies, respectively. Hayashi and Inoue (1991) consider a version of the Q model that allows for multiple capital goods as well as the basic specification, whilst Blundell et al. (1992) estimate a version of the Q model that allows for an $AR(1)$ component in the error term. Both papers report very low coefficients on their Q variables, and find either cash flow or sales terms to be highly significant additional regressors, even allowing these variables to be endogenous and correlated with firm-specific effects.

One potentially important exception to this general pattern of results is provided by Cummins, Hassett and Hubbard (1994), who focus on periods around major tax reforms in the US and report much higher coefficients on their Q variable in these years. This is consistent with their interpretation that major tax reforms provide quasi-experiments that help to identify the effects of economic 'fundamentals' on investment, so that during these periods fluctuations in measured Q are dominated by informative changes in tax parameters rather than uninformative measurement errors. Notice that if this interpretation is correct, the conventional findings discussed in the previous paragraph can only be explained if there are substantial and highly persistent measurement errors in average q that are not easily controlled for by the use of lagged instruments.

Two more recent papers have developed alternative approaches to estimating the Q model which take seriously this possibility of persistent measurement error in average q. Erickson and Whited (2000) consider a GMM estimator based on higher order moment conditions. Their approach can allow for persistent measurement error in the stock market valuation as a measure of the firm's fundamental value, for example as a result of asset price bubbles, provided that the difference between the two values is independent of the fundamental value. They find that additional cash flow variables are no longer significant when this form of measurement error is allowed for in their sample. Bond and Cummins (2001) consider identification of the Q model in the presence of share price bubbles. They show that the parameters may not be identified, using conventional measures of average q, in the more problematic case where the bubble component is

itself correlated with the fundamental value of the firm. To deal with this case, they consider using a direct estimate of the firm's fundamental value, based on forecasts of future profits published by securities' analysts.[49] When using this alternative measure, they find a much higher coefficient on their average q variable than is usually obtained, and neither cash flow nor sales variables are found to be significant.

Gilchrist and Himmelberg (1995) estimate a version of the Abel and Blanchard (1986) model, based on Equation (3.3), for a panel of listed US manufacturing companies. Unfortunately it is not clear whether the model they estimate has a structural interpretation, since in their specification of $(\frac{\partial \Pi_t}{\partial K_t})$ they assume that net revenue Π_t is homogeneous of degree one in K_t, although their specification of adjustment costs depends on current investment and is not homogeneous of degree one in K_t. Certainly in the standard formulation of the Q model, net revenue is homogeneous of degree one in the pair (I_t, K_t), and not homogeneous of degree one in K_t alone.[50] Nonetheless their results are interesting in that they also find a much larger coefficient on their constructed measure of marginal q than on a standard measure of average q, and conditional on this measure of marginal q they also find smaller and weaker coefficients on a cash flow variable for at least some sub-samples of their data. These findings are also suggestive of a potentially severe measurement error problem for conventional measures of average q constructed using stock market valuations.

Results based on the Euler equation approach have been mixed. Unrestricted estimates of investment dynamics have generally been difficult to reconcile with the Euler equation implied by the symmetric, quadratic adjustment costs model, in the sense that estimated coefficients on leads or lags of the investment rate have not implied plausible values for the discount factor (cf. Equation (3.13)). Tests of the over-identifying restrictions implied by the Euler equation have often been rejected, at least for large sub-samples of the data. On the other hand, some papers have reported more reasonable estimates of the structural parameters and non-rejection of the over-identifying restrictions for particular sub-samples, and suggested that these results are consistent with a financing constraints interpretation. See Gilchrist (1991), Whited (1992), Bond and Meghir (1994) and Hubbard, Kashyap and Whited (1995) for examples of this approach.

One conclusion from this literature seems to be that the standard implementations of structural models based on the assumption of symmetric, quadratic costs of adjustment do not provide an adequate characterisation of the observed investment data, at least for a large part of the company data sets that have been used. Recent research in the Q framework suggests that measurement error in stock market valuations, as measures

[49] See also Cummins, Hassett and Oliner (2006) and Bond and Cummins (2000).

[50] Gilchrist and Himmelberg (1995) use the same form for adjustment costs as that in Equation (3.4) – see their Equation (7). Of course the formulation for $(\frac{\partial \Pi_t}{\partial K_t})$ based on the assumption that Π_t is homogeneous of degree one in K_t may provide a good *approximation* if adjustment costs are sufficiently small. Abel and Blanchard (1986) also used this approximation in one of their applications to time series data, but they were careful not to claim that they were identifying a structural adjustment cost function. See their footnote 5.

of the expectations of future profits relevant for investment decisions, should be taken seriously – reflecting the intrinsic difficulty of controlling for *firms'* expectations of future conditions. However it would be too early to conclude that this is the only source of mis-specification. As we emphasised in Section 3.1, there are many candidate explanations for empirical rejections of these models. These results have motivated the huge empirical literature on financing constraints that we consider in the next section, as well as the more recent empirical work on non-convex adjustment costs that we discussed in Section 3.5.

6.2. *Financing constraints*

A major topic of interest in recent microeconometric research on company investment has been to test for the possibility that investment spending is subject to significant financing constraints. In each of the basic investment models outlined above, capital markets were assumed to be perfect in the sense that the firm can raise as much investment finance as it desires at some required rate of return (ρ_{t+1}) that is given exogenously to the firm. In this case the firm's real investment decision is separable from its financial decisions, and investment depends only on the price (i.e. the required rate of return) at which finance is available. Quantitative indicators of the availability of internal finance, such as current profits or cash flow, should affect investment only to the extent that they convey new information about its likely future profitability; and if the maintained structure of the Q model were correct, these financial variables should not appear as significant explanatory variables in an investment model after controlling for a measure of (marginal) q.

This separability between real and financial decisions no longer holds if the firm faces 'imperfect' capital markets, in which internal and external sources of investment finance are not perfect substitutes. We define a firm's investment to be *financially constrained* if a windfall increase in the supply of internal funds (i.e. a change which conveys no new information about the profitability of current investment) results in a higher level of investment spending. Clearly firms are not constrained in this sense in the Q model of Equation (3.2), where given current prices and interest rates, investment depends only on the current and expected future marginal revenue products of capital, as summarised in marginal q through the shadow value of an additional unit of capital. However firms' investment may be financially constrained in 'hierarchy of finance' or 'pecking order' models of corporate finance, in which external sources of finance (for example, from new share issues or borrowing) are assumed to be more expensive than internal sources of finance (for example, from retained earnings).[51]

[51] See, for example, Myers (1984). Notice that this assumption does not require rationing to be present in any of the external capital markets, although it can incorporate rationing [cf. Stiglitz and Weiss (1981)] as a special case in which the cost of external funds becomes infinitely high.

6.2.1. A simple hierarchy of finance model

To illustrate the implications of this assumption rigorously, but as simply as possible, we maintain all the assumptions used to obtain the Q model, except that we introduce an additional cost associated with using external finance. We continue to assume that the firm issues no debt, pays no taxes and is characterised by symmetric information, but we introduce an explicit transaction cost (f_t) per unit of new shares issued. Similar results can be obtained in models where the cost premium reflects asymmetric information or differential taxes, and can be extended to models with debt provided that lending to the firm becomes a risky proposition for lenders beyond some level of debt (i.e. there is some risk of default, all debt is not fully collateralized, and there are 'deadweight' costs associated with defaulting on unsecured debt).[52]

Recognising the distinction between dividends paid (D_t) and the value of new shares issued (N_t), the value of the firm's equity is given by the expected present value of net distributions to shareholders as

$$V_t = E_t\left[\sum_{s=0}^{\infty} \beta_{t+s}(D_{t+s} - N_{t+s})\right] \tag{6.1}$$

whilst the sources and uses of funds identity links dividends and new share issues to the net revenue (Π_t) generated in period t, according to

$$D_t = \Pi_t + (1 - f_t)N_t. \tag{6.2}$$

Introducing non-negativity constraints on dividends and new share issues, with associated shadow values (ν_t^D) and (ν_t^N), the firm's optimisation problem becomes[53]

$$V_t(K_{t-1}) = \left\{ \max_{I_t, L_t, M_t, N_t} \begin{pmatrix} \Pi_t(K_t, L_t, M_t, I_t) - f_t N_t + \nu_t^N N_t \\ \quad + \nu_t^D[\Pi_t(K_t, L_t, M_t, I_t) + (1 - f_t)N_t] \\ \quad + \beta_{t+1} E_t[V_{t+1}(K_t)] \end{pmatrix} \right\}. \tag{6.3}$$

The first-order condition for optimal investment becomes

$$-\left(1 + \nu_t^D\right)\left(\frac{\partial \Pi_t}{\partial I_t}\right) = \lambda_t \tag{6.4}$$

and the Euler equation for λ_t becomes

$$\lambda_t = \left(1 + \nu_t^D\right)\left(\frac{\partial \Pi_t}{\partial K_t}\right) + (1 - \delta)\beta_{t+1} E_t[\lambda_{t+1}]. \tag{6.5}$$

[52] See Hayashi (1985a) and Bond and Meghir (1994) for extensions to models with debt.

[53] Notice that if $f_t = 0$, these non-negativity constraints are redundant. The problem (6.3) reduces to that considered in Sections 2.2 and 3.1, and the firm's financial policy is indeterminate. This is a manifestation of the Modigliani–Miller (1958, 1961) irrelevance theorems.

In addition we now have a first-order condition for optimal new share issues, which for $N_t > 0$ gives

$$\left(1 + \nu_t^D\right) = \frac{1}{1 - f_t} \tag{6.6}$$

and for $D_t > 0$ gives $\nu_t^N = f_t$.

Assuming perfect competition and the same expression for net revenue as in (2.25), the first-order condition for investment becomes

$$\left(\frac{\partial G}{\partial I_t}\right) = \left(\frac{q_t}{1 + \nu_t^D} - 1\right)\frac{p_t^K}{p_t} \tag{6.7}$$

where $q_t = \frac{\lambda_t}{p_t^K}$ is marginal q, as before.

This model has three distinct financial regimes. Retained earnings are the cheapest source of finance, so if the firm has sufficient earnings to finance its desired investment, it will issue no new shares. In this case the non-negativity constraint on dividends is not binding, and the shadow value of an additional unit of internal finance (ν_t^D) is zero. In this regime the basic Q model given by Equation (3.2) describes the firm's investment.

If the firm does not have sufficient earnings to finance its desired investment, the non-negativity constraint on dividends is binding, and the shadow value of internal funds is strictly positive. In this case the firm has to decide whether or not to finance additional investment by using the more expensive external source of finance. If the investment projects that would be foregone by not issuing shares are sufficiently profitable compared to the higher cost of external funds, the firm will choose to issue shares, and using (6.6) its investment in this regime will be described by

$$\left(\frac{\partial G}{\partial I_t}\right) = \left((1 - f_t)q_t - 1\right)\frac{p_t^K}{p_t}. \tag{6.8}$$

However, if the investment projects foregone by not issuing new shares are not sufficiently profitable to warrant paying the higher cost of external funds, the firm will be in a financially constrained position, in which both dividends and new share issues are zero. From the sources and uses of funds condition (6.2) and the net revenue function (2.25), the level of investment expenditure is constrained to the level of cash flow (i.e. $p_t(F_t - G_t) - w_t L_t - p_t^M M_t$). Thus in this constrained regime, windfall changes in cash flow have a direct effect on the level of investment, holding marginal q constant. Allowing the firm to borrow will tend to weaken this sensitivity of investment to windfall fluctuations in cash flow, but will only eliminate it in the special case where debt acts as a perfect substitute for finance from retained earnings.[54]

These results are illustrated in Figure 1, which is adapted from Hayashi (1985a). Investment rates (I/K) are shown on the horizontal axis, and marginal adjustment costs

[54] See Hayashi (1985a) or Bond and Meghir (1994).

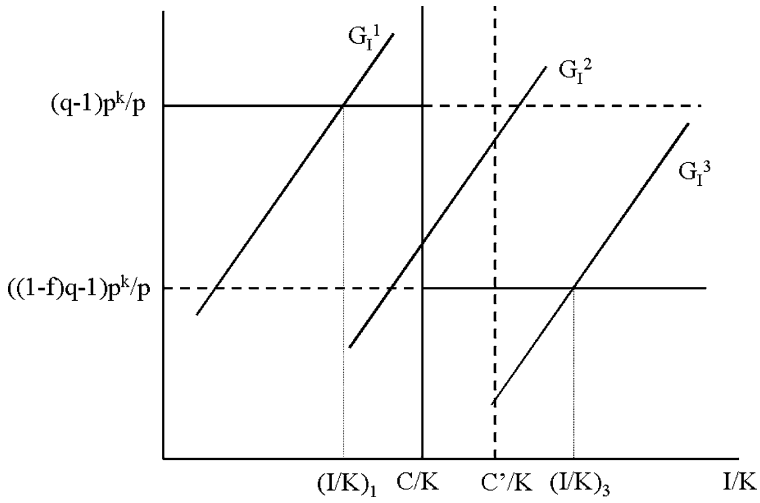

Figure 1. The Q model with financial regimes.

are assumed to be linear in the investment rate, as in (3.4). Values of $(\frac{q_t}{1+v_t^D} - 1)\frac{p_t^K}{p_t}$ are shown on the vertical axis, and the diagram is drawn for a given level of marginal q. The firm can finance investment rates up to (C/K) from internal funds. If the firm with marginal q illustrated has marginal adjustment costs given by the schedule G_I^1, it will choose the investment rate $(I/K)_1$, pay strictly positive dividends and issue no new shares. If the firm has much lower marginal adjustment costs given by the schedule G_I^3, it will choose the investment rate $(I/K)_3$, financed partly (and, at the margin, completely) by issuing new shares, and pay zero dividends. However if the firm faces the intermediate marginal adjustment cost schedule G_I^2, it neither issues shares nor pays dividends, and investment is constrained at the rate (C/K). In this position, a windfall increase in cash flow that allows the firm to finance investment rates up to (C'/K) from internal funds does indeed cause an increase in the firm's investment rate, holding marginal q constant.[55] Whilst we have illustrated these financial regimes by considering different levels of adjustment costs for a given level of marginal q, the same conclusions can be reached by considering different levels of marginal q for a given schedule of marginal adjustment costs.

This model indicates that the simple relationship between investment and marginal q described by Equation (3.2) or (3.5) no longer holds in the presence of financing

[55] Given our timing convention, with current investment immediately productive, this experiment can be thought of as resulting from some exogenous grant or 'helicopter drop' of money to the firm; changes in cash flow resulting from changes to current prices would generally also affect the profitability of current investment. In richer specifications, windfall changes in cash flow could arise from unrelated lines of business [Lamont (1997)] or from certain tax changes [Calomiris and Hubbard (1995)].

constraints. Given the maintained structure of the Q model, a simple test of the null hypothesis that there are no financing constraints can be obtained by including additional financial variables, such as cash flow, on the right-hand side of (3.10). Under the null the estimated coefficients on additional financial variables should be insignificantly different from zero,[56] whilst under the alternative, financial variables will be informative about investment if some firms in the sample are in a financially constrained position. Fazzari, Hubbard and Petersen (1988) and many subsequent papers have exploited this property of the Q model to develop 'excess sensitivity' tests for the importance of financing constraints on firms' investment spending.[57]

It is worth stressing that it is the simple relationship between investment and marginal q that breaks down in the presence of financing constraints, not necessarily the equality between average and marginal q. In the simple model we have described here, we can show that the equality between average and marginal q is maintained, despite the presence of financing constraints, although this need not be the case in richer versions of the hierarchy of finance model. Combining the first-order conditions (6.4) and (6.5) and using the linear homogeneity of the net revenue function, as we did to derive Equation (3.7), gives the expression

$$\begin{aligned}(1-\delta)\lambda_t K_{t-1} &= \left(1+\nu_t^D\right)\Pi_t(K_t, L_t, M_t, I_t) + \beta_{t+1}E_t\left[(1-\delta)\lambda_{t+1}K_t\right] \\ &= E_t\left[\sum_{s=0}^{\infty}\beta_{t+s}\left(1+\nu_{t+s}^D\right)\Pi_{t+s}(K_{t+s}, L_{t+s}, M_{t+s}, I_{t+s})\right] \qquad (6.9)\end{aligned}$$

which depends on current and expected future shadow values of internal funds. However, it is straightforward to show that the relation

$$\left(1+\nu_t^D\right)\Pi_t = D_t - N_t$$

holds in each of the three financial regimes, so using (6.1) we obtain the same expression for marginal q as in Equation (3.9), regardless of which regime the firm is currently in or expects to experience in the future. Thus in this model the basic average Q model (3.10) continues to characterise the investment of those firms that are currently paying positive dividends and issuing no new shares, and in principle the parameters of the adjustment cost function could be identified from this sub-sample. However the literature has focused on testing the null hypothesis that there are no significant financing constraints, for which this equality between average and marginal q under the alternative is not essential.

[56] Although even under the null, current financial variables are likely to be endogenous (e.g. current profits or cash flow would be affected by adjustment cost shocks), and this endogeneity should be allowed for when implementing such tests.

[57] Chirinko (1997) correctly notes that the hierarchy of finance model typically does not imply a linear relationship between investment rates, Q and financial variables, but this is not necessary to motivate these excess sensitivity tests. It is sufficient to show that the simple linear relationship between investment rates and Q is mis-specified under the financing constraints alternative.

6.2.2. *Excess sensitivity tests*

These tests are justified formally because under the maintained structure of the Q model, the Q variable summarises all the information about expected future profitability that is relevant for the current investment decision. This is an important advantage of structural investment models like the Q model or the Euler equation compared to reduced form investment equations in the context of testing for the presence of financing constraints. As we illustrated in Section 3.6, significant effects from financial variables in a reduced form investment equation may simply reflect their role in forecasting future demand or profitability.

Nevertheless this distinction should not be exaggerated, since it relies heavily on all the assumptions that were used to derive the Q model, and becomes blurred once we recognise that these assumptions may be invalid. Thus average Q, for example, would not be a sufficient statistic for investment if the firm operates in imperfectly competitive markets or subject to decreasing returns to scale[58]; standard measures of average Q would not be a sufficient statistic if share prices are subject to bubbles or fads[59]; and the simple linear relation between investment rates and Q would be mis-specified if adjustment costs are not symmetric and quadratic. In all these cases, financial variables may contain additional information that helps to explain investment after controlling for a linear Q term, even under the null hypothesis of no financing constraints.

Fazzari, Hubbard and Petersen (1988) were certainly aware that adding financial variables to the basic Q model is a joint test of all the maintained assumptions of the model, and not simply the assumption of no financing constraints. For this reason, they proposed a test that exploits cross-sectional differences between firms in the relationship between investment and financial variables. The basic idea is that even if the Q model is mis-specified, it may be mis-specified for all firms in a similar way, so that *differences* in the estimated coefficients on additional financial variables in an investment-Q equation may be an indication of *differences* in the impact of financing constraints.

6.2.3. *Sample-splitting tests*

Formally this 'sample-splitting' test can be justified most easily in the following context. Suppose we identify one group of firms for whom the cost premium for external finance is likely to be negligible, and another group of firms for whom the cost premium for external finance may be high. Then if all the assumptions of the Q model were correct apart from the possibility that some firms face financing constraints, we should expect that additional financial variables are insignificant for the first sub-sample; but such financial variables may be significant for the second sub-sample if some of these firms do indeed face binding financing constraints.

[58] See Hayashi (1982), Schiantarelli and Georgoutsos (1990) and Cooper and Ejarque (2003), for example.
[59] See Blanchard, Rhee and Summers (1993), Galeotti and Schiantarelli (1994) and Bond and Cummins (2001), for example.

This sample-splitting test is analogous to that found in the literature which tests for the effect of liquidity constraints on household consumption by investigating heterogeneity in the relationship between consumption and current income across high wealth and low wealth households.[60] Several problems with the test have been noted in the literature. The most favourable outcome would be if we found no evidence of mis-specification for the sub-sample of firms that were considered on *a priori* grounds not to face a cost premium for external finance, and evidence of excess sensitivity to cash flow[61] for the sub-sample of firms that were considered on *a priori* grounds to be potentially subject to financing constraints. Even in this case, it is possible that the result reflects some other source of mis-specification that is only relevant for the second group of firms. For example, Fazzari, Hubbard and Petersen (1988) use the dividend payout ratio as a sample-splitting criterion, arguing that firms facing a high cost premium for external finance will tend to choose a low dividend payout ratio. However it has been noted that their sub-sample of firms with a low dividend payout ratio also tend to be younger and smaller than average, and it may be that share prices are subject to greater pricing errors for this sub-sample. If that were the case, it could explain why average Q is less informative, and additional financial variables more informative, for that sub-sample, even if they are not subject to significant financing constraints.[62]

A second potential problem is that the sample-splitting criterion used may not be exogenous for the investment equations estimated. If the allocation of firms to a particular sub-sample is correlated with shocks to the investment equation, then estimation of the investment model on the endogenously selected sub-sample will be subject to a sample selection bias of the type discussed in Heckman (1979). This suggests that selection criteria based on current financial characteristics or current size may give potentially misleading results unless care is taken to control for the endogeneity of the selection.[63]

An important concern in practice has been the difficulty of finding any sub-samples for which there is no evidence of mis-specification of the basic structural model. The typical finding in studies based on the Q model has been that cash flow variables have significantly larger coefficients in the sub-samples that are considered more likely to be financially constrained, but that such terms also have coefficients that are significantly different from zero for the sub-samples that are considered less likely to be financially constrained. One interpretation is that the Q model is mis-specified for both

[60] See, for example, Hayashi (1985b) and Zeldes (1989).

[61] I.e., significant coefficients on cash flow variables conditional on Q, and controlling for the endogeneity of current cash flow.

[62] This point was noted by James Poterba in his Brookings Panel discussion of Fazzari, Hubbard and Petersen (1988). This particular problem could be avoided by using either an Abel–Blanchard model or Euler equation in place of the average Q model, but the general point remains that the effects of *other* sources of mis-specification of the basic structural model tested may not be common across different sub-samples of firms.

[63] This point was noted by Alan Blinder in his Brookings Panel discussion of Fazzari, Hubbard and Petersen (1988).

sub-samples, but in different ways. Suppose that the Q model is mis-specified even for firms that do not face financing constraints, perhaps because they have market power or non-convex costs of adjustment. This accounts for the significant effect of cash flow even for the sub-sample that is maintained not to be financially constrained. Provided the effects of this mis-specification are similar for the two sub-samples, however, the presence of financing constraints for one group of firms could plausibly explain a significantly higher coefficient for this sub-sample. The difficulty with this interpretation lies in establishing whether the effects of general model mis-specification are indeed similar for the two sub-samples. One possibility would be to investigate directly whether the current (or lagged) financial variables included in the investment equation are more informative predictors of future demand or profitability for one of the two sub-samples; if not, this would cast doubt on one of the leading alternative explanations for their differential importance in the investment equation.[64]

A different interpretation of the same finding is that both sub-samples are subject to financing constraints, but to differing degrees. Thus all firms may face a cost premium for external finance, but some firms may face a much higher cost premium than others. Other things being equal, a bigger transactions cost on new share issues would increase the probability that a firm finds itself in the financially constrained regime in the model we outlined above, and this would tend to increase the sensitivity of investment to fluctuations in cash flow. This interpretation seems reasonable in the simple model of financing constraints we have considered, and many papers have presented evidence of differential cash flow sensitivities as being consistent with the presumption of a higher cost premium for one sub-sample of firms. However there has been some recent controversy about this interpretation, with Kaplan and Zingales (1997) claiming that a higher cost premium for external finance may actually be associated with lower sensitivity of investment to cash flow.[65] To understand this Kaplan–Zingales critique, and its limitations, it is necessary to briefly consider more realistic models of financing constraints in which the firm has access to external finance by issuing debt as well as new shares.

6.2.4. The Kaplan and Zingales critique

A standard model with debt finance, and the one analysed by Kaplan and Zingales (1997), is illustrated in Figure 2.[66] In the absence of financing constraints, this is simply a representation of the first-order condition (2.8) *for a static model of investment*:

[64] Gilchrist and Himmelberg (1995) investigated the forecasting role of cash flow in their study. Notice that this approach can be applied in the context of reduced form investment equations, as well as mis-specified structural models. See Bond, Harhoff and Van Reenen (2003) for an application using error correction models.

[65] See also Fazzari, Hubbard and Petersen (2000) and Kaplan and Zingales (2000).

[66] See, for example, Hubbard (1998). For rigorous treatments of debt finance in the presence of default risk, see Hayashi (1985a) and Bond and Meghir (1994).

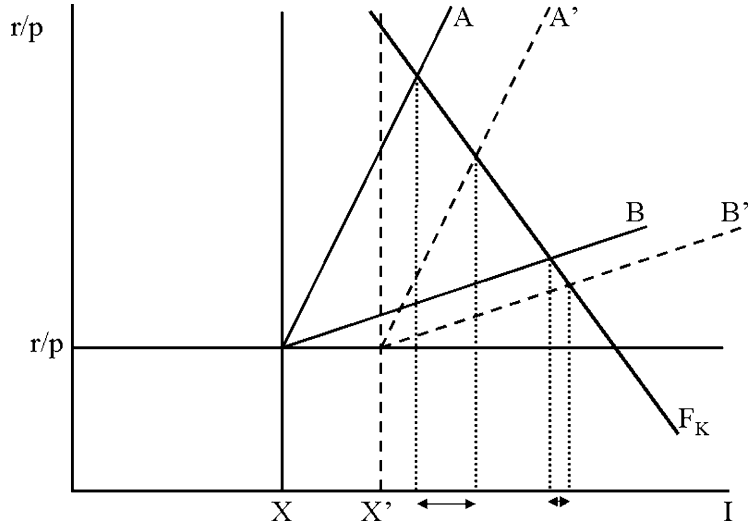

Figure 2. Static model, costly debt, linear MPK.

the downward-sloping line represents the marginal product of capital in the current period,[67] which is equated with the user cost of capital. Given the firm's current cash flow, suppose that borrowing provides a perfect substitute for retained earnings up to the level of investment indicated by X, but becomes increasingly expensive at higher levels of borrowing – perhaps because there is an increasing risk of default and there are deadweight costs associated with bankruptcy. Firms wishing to invest more than X will again find themselves in a financially constrained regime, in which investment is sensitive to windfall fluctuations in cash flow. In this case, a windfall increase in cash flow would increase the level of investment that can be financed without resorting to more expensive debt, say from X to X'. This lowers the marginal cost of external finance for all levels of investment above X, and results in the firm optimally choosing a higher level of investment. Moreover, it appears that firms facing a higher risk premium in the cost of borrowing will display greater sensitivity of investment to cash flow than firms facing a lower risk premium in the cost of borrowing: in Figure 2, a given windfall increase in cash flow has a greater impact on the level of investment for a firm facing the cost schedule A than for a firm facing the cost schedule B.

Kaplan and Zingales (1997) have pointed out that this last conclusion depends on the presumed linearity of the marginal product of capital schedule, as illustrated in Figure 2, and need not hold under alternative assumptions about the production function. To illustrate this possibility, suppose that the marginal product of capital has the

[67] Given the capital stock inherited from the previous period, there is a one-to-one association between the current period's investment level and capital stock.

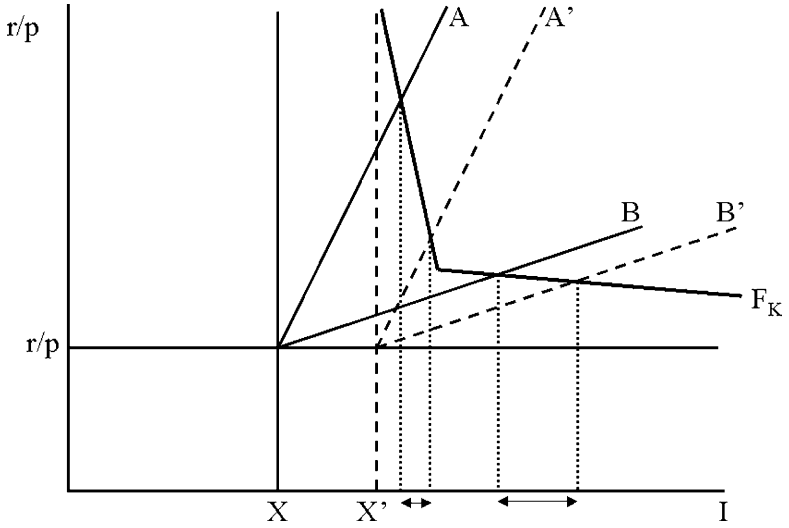

Figure 3. Static model, costly debt, convex MPK.

piecewise linear form shown in Figure 3. In this case, the sensitivity of investment to a given windfall increase in cash flow is greater for a firm facing the cost schedule B than for a firm facing the cost schedule A. More generally, the degree of the financing constraint faced by the firm, as measured by the slope of its cost of external funds schedule, cannot be inferred simply from the sensitivity of investment to cash flow.

This conclusion is certainly correct in the static model analysed by Kaplan and Zingales (1997), but several limitations of the result should be noted. First, the result is obtained under the alternative hypothesis that firms are indeed subject to important financing constraints, and does not undermine the basic excess sensitivity test of the null hypothesis of no financing constraints. As we discuss below, recent papers that have not relied on the average Q model and have controlled for the endogeneity of current cash flow have been more successful in finding some sub-samples of firms for which there is *no evidence* of excess sensitivity to cash flow, whilst the same methods yield significant evidence of excess sensitivity for other sub-samples of the data. Secondly, the Kaplan–Zingales result does not itself provide any alternative explanation for the common empirical finding that there are stronger effects of cash flow on investment for some types of firms than for others; nor does their result rule out the *possibility* that these differential cash flow sensitivities could indeed reflect differences in the severity of financing constraints. Thirdly, their result is obtained in a static model with no adjustment costs, and depends on the curvature of the marginal product of capital schedule. It is not clear that their result extends to a model with strictly convex adjustment costs, in which the first-order condition characterising optimal investment relates marginal adjustment costs to marginal q, as in Equation (3.2), rather than equating the marginal

product of capital to the user cost. In particular, it is not clear that there is a similar result in the model with symmetric, quadratic adjustment costs, in which marginal adjustment costs are a linear function of the investment rate. This limitation is potentially important, since much of the empirical work in this area, including that reported by Kaplan and Zingales (1997) themselves, is based on the assumption of symmetric, quadratic adjustment costs, and not on a static investment model. Finally, their result is obtained in a model where the cost of external funds is increasing at the margin. This may be a reasonable assumption if debt is the only source of external finance. However, in models where the firm can issue new equity as well as debt, and in which the cost of new equity finance is above the cost of internal finance but is not increasing at the margin, the probability of the firm finding itself in the financially constrained regime again depends on the cost premium for using new equity finance – as was the case in the simple model we discussed in Section 6.2.1 above.[68] In these models, a higher sensitivity of investment to cash flow may reflect a higher cost premium for using new equity finance simply because the firm is more likely to be in the financially constrained regime, whether or not investment also responds more to a given change in cash flow within that regime.

In summary, the Kaplan and Zingales (1997) critique is limited to the claim that differing cash flow sensitivities reveal different degrees of financing constraints under the alternative hypothesis that these types of firms are both subject to significant financing constraints; and whilst they present one model in which a greater sensitivity of investment to cash flow could be associated with a lower cost premium for external finance, it is not clear whether this result generalises to models with adjustment costs or a richer specification of the firm's financial policy. Nevertheless, even in the most favourable scenario – for example, in studies which find no excess sensitivity to cash flow for one sub-sample and significant excess sensitivity for another – this sample-splitting test would not establish that heterogeneity in the cost premium for external finance is the correct explanation for this difference. As we stressed in the previous section, it remains quite possible that the different relationship between cash flow and investment among sub-samples of firms can be explained by other sources of mis-specification in the basic structural models that have been used as the basis for these tests.

6.2.5. *Empirical results*

Comprehensive surveys of empirical work investigating the impact of financing constraints on company investment can be found in Hubbard (1998) and Schiantarelli (1996). As we have discussed, much of this literature has followed Fazzari, Hubbard and Petersen (1988) in considering excess sensitivity tests based on adding financial variables to the average Q model, and has tested for differences in the sensitivity of investment to financial variables between sub-samples of firms that are considered more

[68] See Hayashi (1985a), Fazzari, Hubbard and Petersen (1988) and Bond and Meghir (1994) for examples of models of this type.

likely or less likely to be affected by significant financing constraints *a priori*. Sample-splitting criteria that have been considered include dividend payout ratios [Fazzari, Hubbard and Petersen (1988)]; firm size, age or growth [Devereux and Schiantarelli (1990)]; the firm's credit rating [Whited (1992)]; the dispersion in the firm's share ownership [Schaller (1993)]; whether the firm is affiliated to a larger corporate grouping [Hoshi, Kashyap and Scharfstein (1991)]; and whether the firm has a relationship with a particular bank [Elston (1993)]. We do not attempt to review this extensive literature here, but rather focus on a selection of papers that have extended this basic methodology.

Gilchrist and Himmelberg (1995) present results based on a version of the Abel–Blanchard (1986) model, as well as results for an average Q model. They also consider a range of possible sample-splitting criteria, including firm size, dividends and credit ratings. Their results are for a sample of publicly traded US manufacturing companies from Standard & Poor's Compustat database, over the period 1979–1989. They also use a first-differenced GMM estimator of the kind described in Section 4.2 above, to allow for the presence of unobserved firm-specific effects and the endogeneity of both current Q and cash flow variables. Interestingly, when they use their measure of marginal q, they find no significant effects from cash flow for their sub-samples of large firms, and firms with either a bond rating or a commercial paper rating; whilst for their sub-samples of small firms, and firms with no bond or commercial paper ratings, they find significant effects from cash flow conditional on their measure of marginal q (and also conditional on average q). Moreover, Gilchrist and Himmelberg (1995) confirm that these differences in the coefficients on cash flow in their investment equations cannot be accounted for simply by current cash flow being a better predictor of future profitability in some sub-samples than in others; in contrast, they show that the relationship between current cash flow and future cash flow is very similar across the different samples. Although we noted above that their model may not have a structural interpretation, their paper does provide some very interesting evidence on the relationship between investment and cash flow for different groups of US firms.

Bond, Harhoff and Van Reenen (2003) present results based on a reduced form error correction model, for samples of British and German manufacturing companies. They use a system GMM estimator [Blundell and Bond (1998)], which again allows for both firm-specific effects and the endogeneity of their current sales and cash flow variables. They also find no significant effects when cash flow terms are added to their basic error correction specification for investment by German companies, although they do find significant cash flow effects in the same specification for investment by British companies.[69] Again they confirm that this difference does not reflect current cash flow being a better predictor of future cash flow or future sales for the British sample. Bond, Harhoff and Van Reenen (2003) also find that, within their sample of UK companies, there is

[69] Other cross-country comparisons of microeconometric investment equations include Hall et al. (1999) and Bond et al. (2003).

significantly less sensitivity of investment to cash flow for the sub-sample of firms that perform R&D. They suggest that this finding is consistent with a 'deep pockets' explanation for which firms participate in R&D, so that the R&D-performing firms are a self-selected sub-sample where financing constraints associated with the cost of external finance may be less significant.

Bond and Meghir (1994) present a more direct test of the empirical implications of the hierarchy of finance model. They note that this model predicts that the same firm may be financially constrained in some periods and not in others; and that the firm's current dividend and new share issuing behaviour should signal which financial regime the firm is currently in. Thus in Figure 1, for example, the firms in the financially-constrained regime should be paying zero dividends and issuing no new shares; whilst the firms in the unconstrained regimes should be either paying positive dividends or issuing new shares.[70] Using an Euler equation specification for publicly traded UK manufacturing companies, and a GMM estimator that allows for the endogeneity of current financial choices, Bond and Meghir (1994) find that excess sensitivity to financial variables is concentrated in periods when firms pay unusually low dividends (relative to their average payout ratios), and issue no new shares.[71]

Whilst most of the results in this literature appear to be consistent with the possibility of significant financing constraints, at least for some types of firms in some periods, we have emphasized that these tests could also be detecting other sources of mis-specification in the underlying investment models.[72] It should also be noted that there are alternative models in corporate finance, such as Jensen's (1986) 'free cash flow' theory, that could potentially account for the excess sensitivity of investment to cash flow and other financial variables.[73] An important limitation of this literature is that it has not yet developed a convincing structural specification for company investment under the alternative hypothesis that some form of financing constraints is important. This is no easy task, since it would require both a model to allocate firms to different financial regimes, and a rigorous characterisation of optimal financial behaviour in the presence of bankruptcy costs and possibly asymmetric information. Nevertheless these

[70] More generally, whilst most of the sample-splits that have been used in the literature are interpreted as indicating whether or not a particular sub-sample of firms faces a significant cost premium for external finance, those based on current (or average) dividend payout behaviour can be interpreted as indicating whether a particular sub-sample of firms is currently (or predominantly) in a financially-constrained position. The latter tests would continue to have power even if all firms faced potentially significant financing constraints.

[71] Other microeconometric studies based on the Euler equation approach include Gilchrist (1991), Whited (1992) and Hubbard, Kashyap and Whited (1995).

[72] A different possibility, stressed by Gomes (2001), is that cash flow variables may contribute little additional explanatory power in an investment equation, even in the presence of financing constraints. Although accounting for the insignificance of cash flow variables has not been the primary concern of this empirical literature to date, the possibility that these tests have very low power does seem to be present in the simulation model used by Gomes (2001).

[73] In the free cash flow approach, managers have non-value maximising objectives (which may include over-investment) and are subject to less effective monitoring when spending internal funds than external funds.

developments will be important if we are to obtain more compelling evidence in favour of the financing constraints hypothesis, and to obtain useful models for policy simulation analysis in the presence of financing constraints.

6.3. Taxes and the user cost of capital

Compared to the voluminous literature on financing constraints and investment, there has been a dearth of microeconometric studies that focus on estimating the sensitivity of investment to changes in taxes, interest rates or other components of the user cost of capital. This does not reflect any lack of interest in the topic – the relationship between investment and interest rates is crucial to understanding the transmission mechanism of monetary policy, and the effects of taxes and subsidies on investment are crucial for the evaluation of tax policies and potential tax reforms. The frequency of tax changes that are intended to influence investment behaviour suggests that policy makers believe in the effectiveness of this policy, but this view has not received overwhelming support from decades of econometric research based on aggregate time series data.[74]

The limited contribution of microeconometric research to date rather reflects the difficulty of measuring relevant cross-section variation in the user cost of capital. The risk-free nominal interest rate is common to all firms in the same country. Firms producing different products may experience variation in the own-price real interest rate, but measuring this variation requires time series on the prices charged by individual firms which are not widely available. There is more potential for measuring differences in risk-premia across firms, at least to the extent that the relevant risk-premia are well characterised by standard asset pricing models. This is one area where the increasing availability of high-frequency data on stock returns for individual firms may provide a promising direction for future research. Similarly it may be possible to exploit differences across firms in depreciation rates, although the accuracy of measured differences in accounting data may be questionable.

The focus in existing micro studies has generally been on measuring differences in the effects of taxes on the user cost of capital for different firms. One source of this variation is the asymmetry in most corporate tax systems between the treatment of profits and losses. The effective value of tax allowances is reduced for firms in a loss-making or 'tax-exhausted' position by the delay between incurring expenses and being able to claim tax deductions. Unfortunately it is not straightforward to identify which firms are currently in a tax-exhausted position from most publicly available data sources, and it is still harder to forecast when these firms will resume paying taxes. Devereux (1989) and Devereux, Keen and Schiantarelli (1994) have investigated the impact of this type of tax variation on the investment behaviour of UK firms.

[74] See Chirinko (1993a), Hassett and Hubbard (1996) and Caballero (1999) for reviews of this evidence.

Variation also arises because firms use a different mix of capital inputs, some of which receive a more favourable tax treatment than others. Thus firms which tend to use tax-favoured types of capital face a relatively low user cost of capital. Ideally we would want to use investment data disaggregated by type of asset, but even if this is not available it may be possible to exploit the resulting variation across firms or industries in the user cost of capital. Auerbach and Hassett (1992), Cummins, Hassett and Hubbard (1994) and Chirinko, Fazzari and Meyer (1999) have exploited differences in the composition of investment across US industries to measure variation in tax-related components of both tax-adjusted Q variables and the user cost of capital.[75] Cummins, Hassett and Hubbard (1994) estimate significant effects from the cross-section variation that occurs in this measure of the user cost in periods of major tax reforms, with an implied long-run elasticity of the capital stock with respect to the user cost between -0.5 and -1.0 [Hassett and Hubbard (1996)]. Chirinko, Fazzari and Meyer (1999) report statistically significant but smaller estimates, around -0.25, although the estimated returns to scale implied by their reduced form models suggest some doubt about these findings.

Caballero, Engel and Haltiwanger (1995) also exploit this source of tax variation to estimate long-run elasticities of the capital stock with respect to the user cost. Their econometric approach is quite different, in that they rely on co-integration methods to estimate a long-run relationship between the capital-output ratio and the user cost of capital, both in logarithms, as in Equation (2.13) above. This is estimated using time series data for individual US plants in the Longitudinal Research Database, imposing the restriction that the elasticity is equal for all plants within each two-digit industry. Caballero, Engel and Haltiwanger (1995) report estimates ranging from -0.01 to -2.0 across different sectors, with the average being about -1.0. This estimate is similar to that found using co-integration methods on aggregate manufacturing data by Bertola and Caballero (1994) and Caballero (1994).

Whilst it is certainly the case that these micro studies have found evidence consistent with taxes and the user cost of capital having an economically significant effect on capital intensities in the long run, it is perhaps a little early to agree with Hassett and Hubbard (1996) that there is a new 'consensus' on the size and robustness of this effect.

6.4. *Uncertainty*

The relationship between investment and uncertainty has attracted considerable interest in recent theoretical research, and has been investigated in some recent microeconometric studies. Renewed interest in this topic has followed from the development of the literature on 'real options', which stresses that the option of waiting to invest until more information has been revealed can be valuable.[76] If so, then extinguishing this option by investing today should be viewed as a cost. Moreover, the value of this foregone option

[75] See also Cummins, Hassett and Hubbard (1996).

[76] See Dixit and Pindyck (1994) for an excellent introduction to this literature.

will be greater at higher levels of uncertainty, so that current investment may be deterred by a higher level of uncertainty.

This intuitive prediction contrasts with earlier results on the relationship between uncertainty and investment in the context of the Q model. Abel (1983) showed that a higher level of uncertainty would be associated with higher investment in the Q model, although this effect would of course be fully reflected in the behaviour of q, so that in this case there should be no additional effect of uncertainty variables after conditioning on q. Caballero (1991) reconciles these theoretical results by showing that much depends on the nature of the net revenue function and the adjustment costs facing the firm. By maintaining strictly convex adjustment costs and a linear homogeneous net revenue function, the Q model rules out any value associated with the option of delaying investment. In contrast, real options become valuable when there are both non-convex costs of adjustment, such as (partial) irreversibility, and a concave net revenue function. Even then, the effects of a higher level of uncertainty on the average level of the capital stock in the long run are found to be ambiguous [Abel and Eberly (1999), Caballero (1999)], essentially because both investment and disinvestment actions may be deterred by these real option effects. As stressed by Bloom, Bond and Van Reenen (2007), however, a less ambiguous prediction is that a higher level of uncertainty will be associated with slower adjustment of the capital stock, and in particular with a smaller impact effect of demand shocks on current investment.

Empirical work has been limited by the difficulty of finding convincing empirical counterparts to the concepts of 'uncertainty' used in this theoretical literature. Leahy and Whited (1996) considered measures of uncertainty based on the volatility in stock market returns for publicly traded US firms. They found that investment rates were negatively related to these uncertainty measures in simple specifications, but that this effect became insignificant when they conditioned on a standard measure of average q. One concern with this kind of measure is that stock market returns may be subject to 'excess volatility' if share prices are indeed subject to bubbles, fads or other influences over and above firms' fundamental valuations. Nevertheless, Bloom, Bond and Van Reenen (2007) found a smaller impact effect of sales growth on current investment for publicly traded UK firms facing higher volatility in stock returns, although this is in the context of a reduced form, error correction model of investment with no explicit controls for the effect of expected future levels of profitability. A concern in this context is that lower uncertainty (stock market volatility) may be associated with greater optimism about the firm's future prospects. Guiso and Parigi (1999) address both these concerns using data from a specially conducted survey of Italian firms, which asked firms both about expected levels and perceived dispersion in future demand. They also found a weaker effect of expected demand on current investment for firms that perceived greater uncertainty about this future demand, although their results are based on a single cross-section of firms, and the time horizon of these expectational variables is quite short.

Whilst these papers present findings that appear to be consistent with the predicted effects of real options, the development of more convincing measures of uncertainty and more rigorous testing strategies will be required before we can be confident that

these effects are statistically or economically significant. In view of the considerable theoretical and policy interest in this topic, this would seem to be a promising area for further research.[77]

6.5. R&D investment

In modern economies a large proportion of firms' investment is in intangible assets. One of the most important of these are research and development (R&D) expenditures. Economists have long regarded technical change as the most important driver of economic growth, so much of the early microeconometric work on firm level R&D naturally focused on analysing the *effect* of R&D on measures of firm performance (such as productivity, firm market value or patenting activity).[78] R&D is generally cumulated into a 'knowledge stock' and then treated as one type of capital input amongst others. More recently there has been renewed interest in understanding the determinants of firms' R&D decisions, as growth theory refocused on the endogenous decisions of firms to invest in R&D [e.g. Romer (1986), Aghion and Howitt (1992)]. There are relatively few papers which look at the demand for R&D using the dynamic structural models discussed in Section 3,[79] although clearly the results can be extended if we are prepared to treat R&D symmetrically with physical investment. For example, Equation (3.17) would have cross terms in the tangible and R&D capital stocks in addition to Q in the R&D equation.

A key issue here is whether many of the problems inherent with modelling fixed capital are just exacerbated in the case of R&D capital or whether they are qualitatively different. For example, R&D is typically a highly risky and uncertain investment, has large adjustment costs, enjoys many government subsidies (e.g. tax privileges and direct grants), and is subject to strategic gaming (e.g. patent races). But all these issues also arise with fixed capital. Perhaps altogether they add up to a difference in kind, but this is unclear *a priori*.

The main difference between fixed capital and R&D capital is probably in regard to externalities. R&D creates knowledge which is difficult to fully appropriate by the firm making the investment. The 'knowledge spillovers' to other firms create the fundamental public goods problem which gives a rationale for governments to subsidise R&D. Much of the literature on R&D and productivity is motivated by the idea that the social returns to R&D exceed the private returns and there is a large body of work on the empirical search for spillovers [see Griliches (1998) for a summary].

Another problem with comparing R&D to fixed capital is measurement. First, in building up a replacement cost capital stock measure there is typically a benchmark

[77] See Bloom (2006) for a recent structural contribution that incorporates real options effects into investment models.

[78] See, for example, the collection of papers in Griliches (1984).

[79] Hall (1993, 1992), Harhoff (1998), Klette and Moen (1999) report some results for the Euler equation for R&D. Himmelberg and Peterson (1994) have a non-structural version of the Q model.

year when either (i) there was a survey of replacement costs (e.g. fire insurance values) or, more typically, (ii) historic cost stock data can be used. There is no equivalent for R&D so researchers are forced to make assumptions about the pre-sample growth of R&D along with an assumed 'knowledge depreciation' rate. A second problem is that disclosure of R&D in company accounts is far more limited than investment, especially outside the US. Typically, researchers have to deal with the fact that in many countries R&D is a voluntarily disclosed item in company accounts, and therefore subject to serious selectivity biases. Finally, what counts as R&D is less clear than physical investment, as about 90% of reported outlays are current costs and 50% are wages and salaries.

There are large empirical literatures on the effects of firm size, product market structure and labour market institutions on R&D. Since these have been surveyed elsewhere[80] we focus on three issues here – VAR approaches, financial constraints and taxes.

6.5.1. VAR approaches

In light of these modelling and data difficulties, several authors take a vector autoregression (VAR) approach to examining R&D. Lach and Schankerman (1989) focus on unravelling the pattern of Granger causality between R&D and fixed investment by projecting the log of both current R&D and current investment against lagged investment, sales, R&D and other variables. They find that investment does not 'Granger-cause' R&D, but R&D does Granger-cause investment. Although they find corroborating evidence at the industry level in Lach and Rob (1996), others find very different results. Using British data, Nickell and Nicolitsas (1996) find that industry R&D (rather than firm R&D) predicts investment. Toivanen and Stoneman (1998) find the exact opposite result (investment predicts R&D and not *vice versa*). The atheoretical structure of the VAR is problematic here and the interpretation of the correlation pattern (even if it were robust) is difficult. In this context the paper by Pakes (1985) is more satisfactory as his application of dynamic factor demand theory does place more restrictions on the data (a three equation system for R&D, patents and market value). Even though the restrictions are not rejected by the data, it has proved harder to push the theory much further in this direction [e.g. Griliches, Hall and Pakes (1991)], as the framework is fundamentally driven by unobserved stochastic shocks which are only poorly tracked by the observables in the system.

6.5.2. Financing constraints and R&D

A key area of interest for R&D models is the role of financing constraints. It has long been recognised that the asymmetric information problems that lie at the heart of credit constraints may be more important for R&D investments. Uncertainty, lack of collateral and the danger of losing one's ideas to competitors make it likely that firms will rely

[80] See Cohen and Levin (1989) and Menezes-Filho, Ulph and Van Reenen (1998).

on internal sources of finance for R&D more than for other types of investment. On the other hand, the larger adjustment costs for R&D make it unlikely that transitory cash flow shocks will have a very large impact on firms' R&D decisions. Indeed, to the extent that firms only participate in performing any R&D if they can be reasonably sure of not encountering financial constraints, we may see less sensitivity to cash flow by R&D performing firms than non-R&D performing firms.

Hall (1992) analyses a panel of large US manufacturing firms and finds that R&D is significantly correlated with cash flow using a variety of model specifications (reduced form and Euler equations). Himmelberg and Peterson (1994) find evidence that R&D is sensitive to cash flow for small US firms in the high-tech sector. The evidence outside the US is less clear. Hall et al. (1999) use a bivariate VAR approach to examine Granger causality patterns in samples of US, Japanese and French firms in high-tech sectors. They find that the cash flow correlation is far stronger in the US than in the other countries. Mulkay, Hall and Mairesse (2001) also find that the R&D–cash flow correlation is stronger in the US than in France. Bond, Harhoff and Van Reenen (2003) find no effect of cash flow on R&D in their samples of British and German firms. They do find, however, evidence that in Britain cash flow seems to matter for the decision to participate in R&D, whereas it has no effect in the German sample. So the upshot of these studies is that the influence of cash flow on R&D appears stronger in the Anglo-American countries than in Continental Europe or Japan, subject to the concerns we discussed in Section 6.2 above.

6.5.3. Tax-price of R&D

There is a wide variation over time and across countries in the user cost of R&D capital. This is driven by the special treatment of R&D for tax purposes – many countries have tax credits for R&D, super-deductions and accelerated depreciation schedules [see Bloom et al. (1998) for a survey]. Since the tax rules for claiming these benefits often differ depending on corporation tax liabilities, size of firm, current and past R&D spending, region and industry, these tax rules imply that there is a cross-sectional distribution of user costs facing firms in a given year. Hall (1993) and Hines (1994) use US firm panel data to investigate the impact of changes in the user cost on R&D. Dagenais, Mohnen and Therrien (1997) implement a similar methodology using Canadian company data. These authors uncover significant effects of the tax-price on R&D, with a price elasticity of around unity in the long run.[81]

A motivation for these studies is that changes in tax policy may cause some exogenous variation in the price of R&D. Unfortunately, a problem with these studies is that the user cost cannot be taken as truly exogenous as the tax position of individual firms will depend on current shocks which could also influence their R&D decisions. Thus, one still has to use some kind of instrumental variable procedure of the kind discussed in Section 4.2.

[81] Hall (1993) finds larger long-run elasticities than Dagenais, Mohnen and Therrien (1997). Bloom, Griffith and Van Reenen (2002) also find long-run elasticities of around unity using macro data across eight countries.

7. Topics in employment

The demand for labour is a particular case of the general model outlined in Section 2 above. Labour demand is particularly interesting from a policy perspective – the social and political consequences of a 20% fall in the relative demand for less skilled workers will have greater interest than a 20% fall in the relative demand for less sophisticated capital equipment.

A popular approach here is a version of the 'reduced form' models discussed in Section 3.6. For example, analogously to (3.24)

$$l_t = \alpha_1^L l_{t-1} + \beta_0^L l_t^* + \beta_1^L l_{t-1}^*. \tag{7.1}$$

Using (2.14) we have

$$l_t = \alpha_0^L + \alpha_1^L l_{t-1} + \beta_0^L y_t + \beta_1^L y_{t-1} - \beta_0^L \sigma (w-p)_t - \beta_1^L \sigma (w-p)_{t-1} \tag{7.2}$$

where $\alpha_0^L = (1-\alpha_1^L)\sigma \ln a_L (1 - \frac{1}{\eta^D})$. This can also be rewritten (assuming constant returns) in error correction form

$$\begin{aligned}\Delta l_t = {} & \alpha_0^L + \beta_0^L \Delta y_t - \beta_0^L \sigma \Delta (w-p)_t \\ & + \big(1-\alpha_1^L\big)\big[(y-l)_{t-1} - \sigma (w-p)_{t-1}\big].\end{aligned} \tag{7.3}$$

Again, equations of this form can be justified explicitly in a dynamic optimizing framework under quadratic adjustment costs [Nickell (1985, 1986), Bresson, Kramarz and Sevestre (1992)]. Often, researchers (especially in the UK) have assumed a Cobb–Douglas production function and substituted out output for capital. This has the advantage that it is more reasonable to treat capital as predetermined in labour demand equations than output which, in general, must be treated as endogenous. Versions of the Euler equation analogous to (3.11) have also been estimated, although these are less common than in the investment literature.[82]

There are several existing surveys on labour demand. We examine some topics of particular interest arising since the publication of Nickell (1986), Hamermesh (1993) and Hamermesh and Pfann (1996), which give a summary of the literature as it stood at the beginning of the 1990s.[83] The issues of exogenous factor price variation, union bargaining, adjustment dynamics, gross and net flows, and heterogeneity by skill are discussed in turn.

[82] See, for example, Machin, Manning and Meghir (1993), Meghir, Ryan and Van Reenen (1996) and Alonso-Borrego (1998).

[83] We do not attempt to survey the literature which focuses on the adequacy of general equilibrium models of the demand for skills. Interested readers are referred to Heckman and Sedlacek (1985) for an example which examines the selection bias inherent in aggregate studies of wages and labour demand.

7.1. Variation in wages

One important difference between the firm level study of investment and employment is that firm level data often has some information on wages. Typically, these are quite crude measures such as the average wage (wage bill divided by number of workers) but this is a major advantage over capital where variation in the cost of capital between firms has to be constructed by the econometrician as it is absent from firm accounts. Of course, some of the variation in the average firm compensation cost will be variation in the quality mix of workers in the firm (e.g. by skill, gender or ethnic groups) which is conflated with genuine changes in the price of labour facing the firm. Increasingly though, the availability of matched worker-firm panels is enabling researchers to improve their measurement of firm level wage rates.

What are the exogenous sources of variation in the price of labour facing firms? In many contexts changes in the institutional structure surrounding wage determination offer scope for instrumenting firm level wages. Union power, minimum wage changes and regional variation (due to partial labour immobility) offer a much wider range of possibilities than with investment. Unfortunately, when some of these institutionally determined variations in the price of labour are used to examine labour demand, the results have been mixed. The survey in Card and Krueger (1995), for example, illustrates that it is very difficult to ascertain any clear evidence of significant wage elasticities in minimum wage studies.

7.2. Union bargaining

A large sub-literature has developed in estimating employment equations to examine different models of union behaviour. The traditional model of unionization keeps to the neoclassical framework where the firm chooses employment unilaterally. The union, however, has some influence over how the wage is set. The monopoly union model allows the union complete power to set the wage whereas the more general 'right to manage' model allows for genuine bargaining over the wage rate [see Pencavel (1991) or Booth (1995) for an extensive discussion of these models]. These models have the convenient property that the basic structure of the static labour demand model still holds, but the wage will have some firm level variation due to differences in union power between firms.

It is well known that these models are not Pareto efficient and a set-up allowing the firm and union to bargain over both wages and employment can lead to utility gains for both sides [Leontief (1946)]. A second class of 'efficient bargaining' models allows explicitly for such contracts. Ashenfelter and Brown (1986) pointed out that in this case employment will, in general, depend both on the bargained wage and the 'outside' wage (the income received in the event of a breakdown in bargaining). Thus the presence of the outside wage in an employment equation is potentially a test of 'efficient bargaining'. Further generalisations of these models are possible to allow for differential degrees of bargaining power in the wage decision and the employment decision

[Manning (1987)] – this essentially means also including an extra term in union power in the employment equation [Alogoskoufis and Manning (1991)].

There are various criticisms of these approaches. For one the presence of the outside wage in the employment equation could be due to many reasons other than efficient bargaining, such as efficiency wages [Nickell and Wadhwani (1991) attempt to test between the general bargaining model and efficiency wages]. Another criticism is that the testing procedure breaks down once we allow for forward-looking behaviour in the presence of adjustment costs, as the alternative wage can enter the dynamic employment decision rule [Machin, Manning and Meghir (1993)]. Most serious, however, is that it is very difficult to measure what is actually the true alternative wage facing union bargainers. It is quite likely that the average regional or industry wage is an extremely poor measure of this [see MaCurdy and Pencavel (1986)]. This applies even more to measures of union power. The rather inconclusive and fragile results in this literature are likely to stem from this basic problem.[84]

A small, closely related literature seeks to test the adequacy of the neoclassical model of employment determination by analysing behaviour across different ownership structures. Probably the best example of this is Craig and Pencavel (1994) who examine the differences between co-operatives and conventional firms within a single industry – the Pacific Northwest Plywood industry. They found that the standard static model based on profit maximisation was a reasonable description of employment behaviour for the conventional firms, but was quite inadequate for the co-operatives (for example, there was no effect of wages on employment in the latter group).

7.3. Dynamics of adjustment

Asymmetries in adjustment costs have been more of an issue in the labour context because of various regulations aimed at increasing firing costs [e.g. Pfann and Palm (1993), Pfann and Verspagen (1989), Burgess and Dolado (1989)]. Most European countries place a large number of restrictions on the ability of firms to shed labour. Pfann and Verspagen (1989) argued for keeping the assumption of convex adjustment costs but allowing for asymmetries. They suggested an adjustment cost function of the form

$$G(\Delta L) = \frac{1}{2}b(\Delta L)^2 - c\Delta L + e^{c\Delta L} - 1 \tag{7.4}$$

where b and c are parameters in the adjustment cost function. If costs are symmetric then $c = 0$. If the marginal cost of a positive adjustment is greater than the marginal cost of a negative adjustment then $c > 0$. Substitution of this adjustment cost function

[84] See Card (1986), Abowd (1989), Christofides and Oswald (1991) and Boal and Pencavel (1994) for good examples of attempts to examine different models of union behaviour. Another basic problem is that very different models may apply in different industries and bargaining contexts.

into the net revenue function (2.25) leads to a non-linear Euler equation. This can either be estimated directly [Schiantarelli and Sembenelli (1993), Jaramillo, Schiantarelli and Sembenelli (1993)] or with approximation methods [e.g. Pfann and Palm (1993)]. Studies using this approach have found evidence for asymmetries [see Hamermesh and Pfann (1996)].

In an influential paper, Hamermesh (1989) examined monthly data on 7 manufacturing plants. Although the aggregate series appeared smooth, individual plant level employment adjustment was extremely lumpy. Davis and Haltiwanger (1992) produced the first large scale analysis of plant level employment changes in the US. They emphasised the fact that during times of net employment expansion there were very large numbers of firms who were cutting employment, and *vice versa* in recessions. As with Hamermesh (1989), much of the plant level employment changes occurred over short periods.

One branch of the literature has gone on to focus on non-convex adjustment costs as a reason for this heterogeneity (see Section 3.5 above). Caballero, Engel and Haltiwanger (1997) take a similar approach to analysing plant level employment changes as they do for investment (the dataset they use is essentially the same, the Longitudinal Research Database). A crucial issue is how to measure empirically the 'gap' (or labour shortage) term; that is, the difference between the actual level of employment and the target/desired level of employment. They assume that hours per worker can be used to infer a reliable measure of this gap, or rather the difference between hours intensity in the current quarter and the plant's mean hours per worker over the sample period.[85] For a given target level of employment, a large change in employment will lower the size of the gap and therefore the deviation of hours per worker from its long-run mean. They therefore use the coefficient from a regression of the change in employment on the change in hours to build a measure of the gap and then characterise the degree to which a plant actually adjusts towards the optimal level of employment as a function of the size of this gap. They find that the 'adjustment hazard' is not constant as the partial adjustment model would predict but rather increases with the gap. That is, the probability (and proportionate size) of the adjustment increases with the scale of the shortage. They also find two modes of the distribution of employment changes, one at (practically) zero adjustment and another at full adjustment. They conclude that this is consistent with (S, s) types of adjustment behaviour.

One problem with this approach is that the OLS regression of employment changes on hours changes, which is critical in defining the gap term, is subject to endogeneity. Productivity shocks will increase the target level of employment and lead to simultaneous increases in jobs and hours, leading to a strong upwards bias in the relevant parameter. Caballero, Engel and Haltiwanger (1997) attempt to deal with this by conditioning their

[85] An alternative method is to write down explicitly the firm's dynamic optimization problem and use this to calculate the gap. This requires many assumptions about the parameters to calibrate the optimization problem. See Caballero and Engel (1993, 1999) for examples of this approach.

regression sample only on observations where there have been very large changes in both employment and hours. They argue that in these periods, the changes in employment targets will be swamped by the effects of large changes in hours and employment. Cooper and Willis (2004) suggest that this approach may be misleading by analysing some simulated data generated from a purely convex adjustment cost model in a stochastic environment. In such an environment, the only periods of large employment and hours changes are exactly those in which there are large changes in the employment target levels. Cooper and Willis (2004) also suggest that (mis)measuring the gap using the Caballero, Engel and Haltiwanger (1997) methodology on their simulated data produces the kind of non-linearities in the aggregate data that Caballero et al. interpreted as evidence of non-convexities. However their simulation results are based on samples that cover more time periods than those used in empirical work, and the degree to which mis-specification of the gap can result in the appearance of quantitatively significant non-linearities remains controversial.

Cooper and Willis (2004) argue for a more explicit structural approach to address the issues of non-convex adjustment costs. An example of this is contained in Rota (1998) who uses data on Italian firms. These firms are on average smaller than the plants studied in Caballero, Engel and Haltiwanger (1997). There is a large mode at (absolutely) zero adjustment and Rota (1998) uses this to define three regimes (adjust up, adjust down and do not adjust). The adjustment regimes are characterised by an Euler equation analogous to that described in Section 3.3 above and the selection rule into regimes is determined by an ordered probit. Estimation of the Euler equations requires using the estimated parameters from the ordered probit to correct for the endogenous selection into adjustment regimes. Apart from the usual problems discussed above, her study raises several additional issues. One worry is how the regime selection rule is separately identified from the employment rule [this problem also arises in Hamermesh (1989)]. A second (and related) issue is the fact that the model is silent on the structural form of the selection rule which determines when an adjustment takes place. An attempt to explicitly implement a dynamic discrete model of employment adjustment is Aguirregabiria and Alonso-Borrego (1999),[86] who consider a model with linear (but not fixed) adjustment costs. The adjustment costs are amplified by the firing costs due to Spanish labour regulations and the authors examine the impact of a reduction in these for workers on temporary contracts. They estimate the productivity shock through a first stage production function and then use this explicitly in the dynamic optimization problem for the employment decision in a second stage. They found important effects of the reform in improving employment and job mobility. Although this approach is an advance it does hinge critically on the correct specification and estimation of a production function to identify the technology shock.[87]

[86] See also Aguirregabiria (1997).

[87] For discussions of some of the many problems with estimating production functions with firm level data, see Griliches and Mairesse (1998), Olley and Pakes (1996) and Blundell and Bond (2000).

7.4. Gross vs. net flows

Hamermesh and Pfann (1996) emphasize that most studies of adjustment costs consider net rather than gross employment flows. A firm might have no net change in employment but hire 1000 workers and sack 1000 workers. It is likely that this firm will bear more adjustment costs than one which had no gross changes in employment at all.[88]

Information on turnover is rare in most micro datasets, but a few studies have started to examine the issue in more detail. Hamermesh and Pfann (1996) analysed a large sample of Dutch plants and decomposed aggregate gross flows. They found that firm level gross flows accounted for a substantial proportion of the total. Abowd, Kramarz and Margolis (1999) examined French data on the entry and exit of workers from firms and attempted to use information on the size of costs associated with the movements of workers. They found that there were very high fixed costs associated with firing workers and most adjustment was through varying the hiring rate.

7.5. Heterogeneous labour and skill-biased technical change

Most firm level datasets only have information on total employment. Aggregation over different types of labour can cause many problems, for example the appearance of spurious dynamics [Nickell (1986)]. If there is access to data disaggregated by skill type, for example, a whole range of important questions are opened up. Most obviously there is the question of how the wage elasticity of labour demand differs between different groups of workers. Bresson, Kramarz and Sevestre (1992), for example, estimated employment equations for 3 different types of labour in 586 French manufacturing firms – they found that the wage elasticities were greatest for the least skilled workers.

We have discussed the issue of multiple quasi-fixed factors in Section 3.4 above and the analysis extends in a straightforward manner to multiple types of labour. Many authors have been interested in how adjustment costs might differ for different groups of workers: for example, are they higher for skilled than for unskilled workers? The issue of dynamic complementarity and substitutability has been considered [Nadiri and Rosen (1969)].

The debate over the (static) elasticities of complementarity and substitution between heterogeneous labour, capital and technology has been a topic of long standing interest to economists. Discussions have been enlivened in recent years by the rapid increase in the college-high school premium in the United States, Great Britain and many other countries. Many researchers have argued that this is primarily due to skill-biased technological change. We do not intend to review all the arguments here [see Autor and Katz (1999) for a survey] but we will focus on a more narrow set of questions. First we

[88] Note that this is a different meaning of gross job flows than that found in the job creation and destruction literature discussed above.

consider some of the meanings of technology-skill complementarity. Second, we critically consider methodologies for testing its size and existence. Finally we review some of the empirical results, particularly those based on micro data.

We examine the neoclassical analysis of skill-biased technical change initially in the context of the static factor demand model described in Section 2.1. A closely related issue is capital-skill complementarity.[89] At issue here is the Hicks–Allen partial elasticity of substitution discussed above. One could clearly take the same approach to technology, viewing it as a form of (partially appropriable) 'knowledge capital' [Griliches (1998)] and modelling it as simply another form of capital input. The alternative approach is more traditional, treating technology as a free good available to all firms in the economy. Note that the key difference is that in the first approach technology is essentially a choice variable for the firm: one factor among others. It may have special features (such as non-rivalry as emphasized in the endogenous growth literature) but can be considered as a choice variable for the firm. In the second approach technology is 'manna from heaven'. It is exogenous and does not change with economic conditions (or at least is treated as such).[90] Both these notions are discussed in the following simple model.

Consider a production function for value-added (Y) with three factors (skilled labour (S), unskilled labour (U) and capital (K)). Using the results from Section 2.1.1 the share equations derived from the translog cost function (Equation (2.15)) are

$$\begin{aligned} S_S &= \alpha_S + \gamma_{SS} \ln W_S + \gamma_{SU} \ln W_U + \gamma_{SK} \ln W_K + \gamma_{SY} \ln Y + \phi_{\tau W_S} t, \\ S_U &= \alpha_U + \gamma_{US} \ln W_S + \gamma_{UU} \ln W_U + \gamma_{UK} \ln W_K + \gamma_{UY} \ln Y + \phi_{\tau W_U} t, \\ S_K &= \alpha_K + \gamma_{KS} \ln W_S + \gamma_{KU} \ln W_U + \gamma_{KK} \ln W_K + \gamma_{KY} \ln Y + \phi_{\tau W_K} t. \end{aligned} \tag{7.5}$$

The restrictions that can be placed on the parameters of the share equations are as follows. Symmetry will mean $\gamma_{ij} = \gamma_{ji}$. Homogeneity means we also have $\sum_{i=S,U,K} \gamma_{ij} = 0$ for all j factors, $\sum_{i=S,U,K} \gamma_{iY} = 0$, $\sum_{i=S,U,K} \phi_{\tau W_i} = 0$ and $\sum_{i=S,U,K} \alpha_i = 1$. Coupled with the fact that the shares add up to unity, one equation becomes redundant and we need only estimate the system

$$\begin{aligned} S_S &= \alpha_S + \gamma_{SS} \ln(W_S/W_U) + \gamma_{SK} \ln(W_K/W_U) + \gamma_{SY} \ln Y + \phi_{\tau W_S} t, \\ S_K &= \alpha_K + \gamma_{KS} \ln(W_S/W_U) + \gamma_{KK} \ln(W_K/W_U) + \gamma_{KY} \ln Y + \phi_{\tau W_K} t. \end{aligned} \tag{7.6}$$

Constant returns to scale implies the restriction that $\gamma_{SY} = \gamma_{KY} = 0$. Symmetry ($\gamma_{SK} = \gamma_{KS}$) implies a further cross equation restriction. Given estimates of the parameters we can calculate all the elasticities of substitution from the formulae in Section 2.1.1.

[89] Griliches (1969) is the pioneering paper here finding evidence in favour. Weiss (1977), by contrast, did not find consistent evidence across sectors using more disaggregated data by skill type.

[90] The exogenous technology approach is where the term 'skill-biased technical change' originated from. The endogenous technology approach is closer to the idea of complementarity proper.

A positive coefficient γ_{SK} implies substitutability, but complementarity between two factors requires not only that the coefficient γ_{SK} be negative, but also more negative (in absolute value) than the product of the factor shares.[91] The bias of technical change depends on the values of the $\phi_{\tau W}$ parameters. The price elasticities are $S_j\sigma_{ij}$ and are also easily calculated from the estimated parameters (and the predicted shares).

There are various ways to bring dynamic considerations into this equation. A simple way is to treat capital as quasi-fixed.[92] Thus, instead of the long-run cost function many researchers would consider a short-run variable cost function [e.g. Brown and Christensen (1981)]. This recognises that the quasi-fixed factors are not at their long-run optimal values (without being explicit over the adjustment dynamics). In comparison to Equation (2.15) we replace the cost of capital with the quantity of capital

$$\begin{aligned}\ln VC = {} & \ln\alpha_0 + \sum_{i=1}^{n}\alpha_i \ln W_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}\ln W_i \ln W_j \\ & + \alpha_Y \ln Y + \frac{1}{2}\gamma_{YY}(\ln Y)^2 + \sum_{i=1}^{n}\gamma_{iY}\ln W_i \ln Y + \gamma_{KY}\ln K \ln Y \\ & + \gamma_K \ln K + \frac{1}{2}\gamma_{KK}(\ln K)^2 + \sum_{i=1}^{n}\gamma_{iK}\ln W_i \ln K \\ & + \phi_\tau t + \frac{1}{2}\phi_{\tau\tau}t^2 + \phi_{\tau Y}t\ln Y + \sum_{i=1}^{n}\phi_{\tau W_i}t\ln W_i + \phi_{\tau K}t\ln K \end{aligned} \tag{7.7}$$

where $i = S, U$ $(n = 2)$. Using the same logic as before, using Shephard's (1953) Lemma for the variable factors and imposing homogeneity and symmetry, we end up with a single (variable cost) share equation

$$S_S = \alpha_S + \gamma_{SS}\ln(W_S/W_U) + \gamma_{SK}\ln K + \gamma_{SY}\ln Y + \phi_{\tau W_S}t \tag{7.8}$$

where S_S is now the share of skilled workers in variable costs (the wage bill). If we want to impose constant returns here $\gamma_{SK} = -\gamma_{SY}$. Assuming that this is true, we can write the wage bill share equation as

$$S_S = \alpha_S + \phi_{\tau W_S}t + \gamma_{SS}\ln(W_S/W_U) + \gamma_{SK}\ln(K/Y). \tag{7.9}$$

As before, the Allen partial elasticity of substitution between skilled and unskilled labour is greater (less) than unity as $\gamma_{SS} > (<)\,0$. The coefficient on the capital intensity variable should be positive to be consistent with capital-skill complementarity. Explicit calculation of the size of the elasticity of substitution/complementarity requires

[91] See Equation (2.21). Even if all factors were substitutes, one might still be interested in whether the elasticity of substitution with capital was greater for unskilled workers than for skilled workers.

[92] This also has the practical advantage that, as we discussed in Section 6.3, measuring exogenous variation in the user cost of capital is extremely difficult.

additional information, such as direct estimates of the cost function parameters. If the sign of $\phi_{\tau W_S}$ (essentially the time trend) is positive, this is consistent with skill-biased technical change.

There are two major problems with using this method as a way for examining skill-biased technical change. First, the time trend could be picking up many other aspects of the economic environment rather than just technical change. This is the standard problem of treating technology as a residual. A natural response to this is to find more direct proxies for technology. Clearly all the usual problems arise in that there is no perfect measure of technical change, but some observable measure (R&D, patents, etc.) seems preferable to assuming the residual trend is all technology. Once we do use explicit measures of technical change, however, we run into the second problem that firms have influence over technical progress. R&D for example, as discussed in Section 6.5, is also a choice variable. One could set the problem up as a model where we consider two capital stocks, knowledge capital (G) and physical capital (K), in the variable cost function.[93] This would imply adding an extra term $\ln(G/Y)$ to (7.9), giving[94]

$$S_S = \alpha_S + \phi_{\tau W_S} t + \gamma_{SS} \ln(W_S/W_U) + \gamma_{SK} \ln(K/Y) + \gamma_{SG} \ln(G/Y). \tag{7.10}$$

We have not discussed the method of including technology variables in earnings equations as a way of examining skill biases [e.g. Krueger (1993)]. The omission is deliberate [see Chennells and Van Reenen (1997) for a longer discussion]. The theoretical basis of such an equation is unclear. In the neoclassical model technology shocks can affect the demand for labour, but the wage is exogenous to the firm as it is determined in the aggregate labour market (this is implicit in the structure discussed above). Under this view the significance of technology indicators in individual earnings equations is likely to capture unobserved ability of individuals which is correlated with both wages and the use of new technologies. There is evidence that this is indeed the case [e.g. DiNardo and Pischke (1997)].

Although it uses industry data, a key paper in this area is Berman, Bound and Griliches (1994), who estimate versions of Equations (7.9) and (7.10) on 4 digit US manufacturing data in long differences.[95] They use R&D expenditures and computer investment as their measures of technical change. These proxies for technology are found to have a positive and significant association with the growth in the wage bill share of non-production workers, the computer variable accounting for about a third of

[93] The distinction between skill-biased technical change and capital-skill complementarity can become murky. In the standard approach technology is exogenous and capital is chosen by the firm. Capital-skill complementarity is a conjecture about the shape of isoquants holding output constant. Technical change, however, causes a move to a new isoquant. Embodied technical change makes the distinction even less clear.

[94] Calculation of the elasticities of substitution/complementarity with two quasi-fixed factors is complex. The sign of the cross-elasticity will depend on both γ_{SK} and γ_{SG} [see Brown and Christensen (1981)].

[95] Although (like much of the subsequent literature) they replace wages with time dummies due to the problem that industry wage changes reflect a mix of genuine change in the price of labour and changes in the skill mix.

the increase in the share. Autor, Katz and Krueger (1998) extend this study over a longer time period (from the 1940s to early 1990s) and to non-manufacturing industries. They corroborated the importance of technical change (especially computer use) in accounting for the increase in skilled workers as a proportion of the wage bill. Machin and Van Reenen (1998) compare the US to 6 other countries (Denmark, France, Germany, Japan, Sweden and the UK). They find results which broadly support the importance of skill bias using their measure of R&D intensity. Other papers with country-specific analyses have also tended to find evidence of skill-biased technical change [e.g. Hansson (1997) for Sweden and Gera, Gu and Lin (2001) for Canada], but Goux and Maurin (2000) are more sceptical about its importance in France.

Aggregation may be a serious problem for these industry studies. The Longitudinal Research Database (LRD), a manufacturing panel dataset for the population of larger plants, has been a prime resource in the USA. Doms, Dunne and Troske (1997) and Dunne, Haltiwanger and Troske (1997) both find evidence of skill bias, but Doms, Dunne and Troske (1997) stress that they cannot find evidence for significant effects in the time series dimension of their data. This is a worrying result, because it does suggest that some other unmeasured factor may be driving both skills and technology. On the other hand, measurement error issues and the fact that they use counts of production technologies (rather than computer usage) might account for their results. Indeed, when they use measures of computer capital instead of the count-based measure they find evidence of significant skill bias even in the time series dimension. Adams (1999) focuses on firms mainly operating in the chemical industry. He finds that firm R&D in the same product field as that produced by the plant is associated with skill bias.

Duguet and Greenan (1997) use an innovations survey to estimate cost share equations for a panel of French manufacturing firms, 1986–1991, in long differences. They find evidence for skill bias and argue that it comes primarily from the introduction of new products, although their results here are mixed. One problem with subjective innovations surveys is the comparability of the notion of innovation across different firms. An interesting extension, given the increasing availability of this type of innovation survey, would be to use the longitudinal aspect of the panel when the innovation questions are asked to the same firms in future. Machin (1996) uses the British Workplace Industrial Relations Survey (WIRS) panel 1984–1990, which contains information on the introduction of computers and also finds evidence for skill bias. Haskel and Heden (1999) use data on about 10,000 British manufacturing establishments from the Annual Respondents Database (ARD) panel to estimate (7.10). They also find that changes in the wage bill share of non-production workers are correlated significantly with the intensity of investment in computer technology. Aguirregabiria and Alonso-Borrego (2001) use firm level Spanish panel data and attempt to control for some of the non-convexities discussed above. They find effects from the first introduction of 'technological capital', but they find no effects from subsequent increases in the stock of this capital or from R&D.

Taken as a whole we draw three conclusions from this body of empirical work. First, there does appear to be considerable support for the notion of skill-biased technical

change across a range of studies, and these are usually (but not always) robust to controlling for fixed effects. Secondly, there have been few attempts to find instrumental variables to deal with the potential endogeneity of technology. Candidate instruments could include government-induced schemes to alter the incentives to accumulate technological capital (such as R&D tax credits, government grants, etc.). Thirdly, there are surprisingly few studies which try to analyse the mechanisms by which technological change translates into higher demand for skills[96]. One mechanism is through organisational changes such as delayering, decentralisation and giving greater autonomy to workers. These organisational factors have been found to be important in the case study evidence and in the literature on the productivity paradox (investigating why computers have not raised measured productivity by as much as might have been expected). Some preliminary work suggests that this organisational restructuring could be the link between technology and labour demand [Bresnahan, Brynjolfsson and Hitt (2002), Caroli and Van Reenen (2001)].

8. Conclusions

We can summarise the main themes from this chapter as follows. Structural microeconometric models of investment and employment are useful for testing hypotheses about the environment in which firms make decisions about their factor inputs: is investment spending subject to financing constraints? Are employment levels subject to union bargaining? Given that complete adjustment of capital stock and employment levels does not occur immediately, these structural models need to be dynamic in nature. This presents a major challenge for econometric modelling, since current decisions depend on unobserved expectations of future demand conditions and factor prices.

Structural dynamic models developed in the 1980s generally assumed that firms face strictly convex costs of adjustment. These models rationalise slow adjustment, and allow structural econometric specifications to be derived that control for the role of unobserved expectations. Examples include the Q model of investment and the Euler equation models that have been used in both the investment and employment literatures. However, these models predict a smooth, gradual pattern of adjustment. Recent work, particularly that which uses plant level data, has suggested that a pattern of infrequent, large adjustments may be more relevant for both capital and labour. Moreover, the structural dynamic models based on strictly convex adjustment costs have generally been rejected in microeconometric tests. Initial optimism that the rejection of these models with more aggregated data may be attributed to aggregation biases does not appear to have been well founded.

Whilst there is now a broad consensus that these traditional structural models appear to be inadequate, there is less agreement on which sources of potential mis-specification

[96] For a recent exception see Autor, Levy and Murnane (2003) who argue that IT substitutes for both manual and non-manual 'routine' tasks.

are most important. Does this reflect the importance of capital market imperfections, or non-convex components of adjustment costs, or something else? How important is the measurement error introduced by different approaches to controlling for firms' expectations of future conditions, particularly when stock market valuations are used? How important are the simplifying assumptions typically made concerning market structure? Whilst a lot of research in the last decade has implemented rigorous tests of relatively simple structural specifications against specific alternatives, surprisingly little progress has been made in developing richer structural models that incorporate these features. This balance will need to be redressed if we are to provide more convincing evidence that particular features of the firm's technology and environment are important for understanding investment and employment behaviour.

As in other areas of microeconometric research, such as household consumption and labour supply behaviour, it is important to recognise that all aggregation problems are not circumvented by the use of data on microeconomic units. Annual investment spending by a large, publicly traded company is clearly aggregated both over time and over different types of capital goods, and may be aggregated over multiple plants or subsidiary firms. Total employment is also aggregated over heterogeneous types of workers, and a constant level of employment may disguise significant inflows and outflows. Identification of structural models of investment and employment dynamics may require more serious attention to be paid to these aggregation issues than has generally been the case in previous research.

Another striking feature of this literature is the limited attention that has been paid to directly estimating long-run price elasticities of demand for capital and labour inputs. This is largely explained by the paucity of microeconomic data on factor prices. New data sources may allow significant progress to be made in this area. In the employment context, the development of matched panels covering both individual workers and individual firms should provide more accurate information on wage rates paid by individual firms than has been available hitherto. On the investment side, high frequency data on stock returns may allow cross-firm variation in the risk premium component of the user cost of capital to be exploited.

Another major area of interest that merits further research is the impact of technological and organisational change on the demand for capital and labour. On the employment side, there has been considerable research into the impact of skill-biased technical change, but surprisingly little micro research has addressed the relationship between technological opportunities and investment. Given the enormous policy interest in the effects of technical progress, additional work is required on the nature of these effects and the transmission mechanisms through which technical change affects investment and employment.

Finally we note that whilst the vast majority of microeconometric research has used data for firms and plants in the manufacturing sector, manufacturing industry now accounts for a comparatively small and declining share of aggregate investment, employment and GDP in most developed economies. Similarly the globalization of business activities has meant that multinational corporations now account for a significant and

growing share of total domestic investment and employment in many countries. Multinational firms have opportunities to substitute between domestic and foreign factors of production, which may make their investment and employment behaviour qualitatively different from that of purely domestic firms. Greater emphasis on the behaviour of multinational companies and firms in service sectors is likely to be required if these microeconometric studies are to provide useful insights into the broader behaviour of investment and employment.

References

Abel, A.B. (1980). "Empirical investment equations: An integrative framework". In: Brunner, K., Meltzer, A. (Eds.), On the State of Macroeconomics. In: Carnegie–Rochester Conference Series, vol. 12, pp. 39–93.

Abel, A.B. (1983). "Optimal investment under uncertainty". American Economic Review 73, 228–233.

Abel, A.B., Blanchard, O.J. (1986). "The present value of profits and cyclical movements in investment". Econometrica 54, 249–273.

Abel, A.B., Eberly, J.C. (1994). "A unified model of investment under uncertainty". American Economic Review 84, 1369–1384.

Abel, A.B., Eberly, J.C. (1996). "Investment and q with fixed costs: An empirical analysis". Mimeo. The Wharton School, University of Pennsylvania.

Abel, A.B., Eberly, J.C. (1999). "The effects of irreversibility and uncertainty on capital accumulation". Journal of Monetary Economics 44, 339–377.

Abowd, J.M. (1989). "The effect of wage bargains on the stock market value of the firm". American Economic Review 79 (4), 774–809.

Abowd, J.M., Kramarz, F. (1999). "The analysis of labor markets using matched employer-employee data". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3B. North-Holland, Amsterdam.

Abowd, J.M., Kramarz, F., Margolis, D.N. (1999). "High wage workers and high wage firms". Econometrica 67 (2), 251–333.

Adams, J. (1999). "The structure of firm R&D and the factor intensity of production". Review of Economics and Statistics 81, 499–510.

Aghion, P., Howitt, P. (1992). "A model of growth through creative destruction". Econometrica 60, 323–351.

Aguirregabiria, V. (1997). "Estimation of dynamic programming models with censored dependent variables". Investigaciones Economicas XXI (2), 167–208.

Aguirregabiria, V., Alonso-Borrego, C. (1999). "Labor contracts and flexibility: Evidence from a labor market reform in Spain". Universidad Carlos III de Madrid Working Paper No. 99-27.

Aguirregabiria, V., Alonso-Borrego, C. (2001). "Occupational structure, technological innovation and reorganization of production". Labour Economics 8 (1), 43–73.

Ahn, S.C., Schmidt, P. (1995). "Efficient estimation of models for dynamic panel data". Journal of Econometrics 68, 5–28.

Akerlof, G. (1970). "The market for lemons: Quality uncertainty and the market mechanism". Quarterly Journal of Economics 84, 488–500.

Allen, R. (1938). Mathematical Analysis for Economists. Macmillan, London.

Alogoskoufis, G., Manning, A. (1991). "Tests of alternative wage-employment bargaining models with an application to the UK aggregate labour market". European Economic Review 35, 23–37.

Alonso-Borrego, C. (1998). "Demand for labour inputs and adjustment costs: Evidence from Spanish manufacturing firms". Labour Economics 5, 475–497.

Alonso-Borrego, C., Arellano, M. (1999). "Symmetrically normalised instrumental-variable estimation using panel data". Journal of Business and Economic Statistics 17, 36–49.

Altug, S., Miller, R.A. (1990). "Household choices in equilibrium". Econometrica 58 (3), 543–570.

Altug, S., Miller, R.A. (1991). "Human capital, aggregate shocks and panel data estimation". Institute for Empirical Macroeconomics, Discussion Paper No. 47. Federal Reserve Bank of Minneapolis.

Anti Nilsen, O., Schiantarelli, F. (2003). "Zeroes and lumps in investment: Empirical evidence on irreversibilities and non-convexities". Review of Economics and Statistics 85 (4), 1021–1037.

Arellano, M., Bond, S.R. (1991). "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations". Review of Economic Studies 58, 277–297.

Arellano, M., Bover, O. (1995). "Another look at the instrumental-variable estimation of error component models". Journal of Econometrics 68, 29–52.

Arellano, M., Honoré, B. (2001). "Panel data models: Some recent developments". In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics, vol. 5. North-Holland, Amsterdam.

Arrow, K.J. (1968). "Optimal capital policy with irreversible investment". In: Wolfe, J.N. (Ed.), Value, Capital and Growth: Papers in Honour of Sir John Hicks. Edinburgh University Press.

Arrow, K.J., Chenery, H.B., Minhas, B., Solow, R.M. (1961). "Capital-labor substitution and economic efficiency". Review of Economics and Statistics 43, 225–254.

Ashenfelter, O., Brown, J. (1986). "Testing the efficiency of employment contracts". Journal of Political Economy 94 (3), S40–S87.

Auerbach, A.J., Hassett, K.A. (1992). "Tax policy and business fixed investment in the United States". Journal of Public Economics 47, 141–170.

Autor, D., Katz, L. (1999). "Changes in wage inequality". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3A. North-Holland, Amsterdam.

Autor, D., Katz, L., Krueger, A. (1998). "Computing inequality: Have computers changed the labor market?". Quarterly Journal of Economics 113, 1169–1214.

Autor, D., Levy, R., Murnane, D. (2003). "The skill content of recent technological change: An empirical exploration". Quarterly Journal of Economics 118 (4), 1279–1333.

Barnett, S.A., Sakellaris, P. (1998). "Nonlinear response of firm investment to Q: Testing a model of convex and non-convex adjustment costs". Journal of Monetary Economics 42, 261–288.

Bean, C.R. (1981). "An econometric model of manufacturing investment in the UK". Economic Journal 91, 106–121.

Berman, E., Bound, J., Griliches, Z. (1994). "Changes in the demand for skilled labor within US manufacturing: Evidence from the annual survey of manufacturers". Quarterly Journal of Economics 109, 367–398.

Berndt, E. (1991). The Practice of Econometrics. Addison–Wesley, Reading, MA.

Bertola, G., Caballero, R.J. (1994). "Irreversibility and aggregate investment". Review of Economic Studies 61, 223–246.

Blanchard, O.J., Watson, M. (1982). "Bubbles, rational expectations and financial markets". In: Wachtel, P. (Ed.), Crises in the Economic and Financial Structure. Lexington, MA.

Blanchard, O.J., Rhee, C., Summers, L.H. (1993). "The stock market, profit and investment". Quarterly Journal of Economics CVIII, 115–134.

Bloom, N. (2006). "The impact of uncertainty shocks: Firm level estimation and a 9/11 simulation". Centre for Economic Performance Discussion Paper No. 718.

Bloom, N., Chennells, L., Griffith, R., Van Reenen, J. (1998). "The effects of tax treatment on the changing cost of R&D: Evidence from eight countries". In: Lawton-Smith, H. (Ed.), The Economics of Regulation of High Technology Industries. Oxford University Press, Oxford.

Bloom, N., Griffith, R., Van Reenen, J. (2002). "Do R&D tax credits work: Evidence from an international panel of countries 1979–1997". Journal of Public Economics 85, 1–31.

Bloom, N., Bond, S.R., Van Reenen, J. (2007). "Uncertainty and investment dynamics". Review of Economic Studies 74 (2), 391–415.

Blundell, R.W., Bond, S.R. (1998). "Initial conditions and moment restrictions in dynamic panel data models". Journal of Econometrics 87, 115–143.

Blundell, R.W., Bond, S.R. (2000). "GMM estimation with persistent panel data: An application to production functions". Econometric Reviews 19, 321–340.

Blundell, R.W., Stoker, T. (2007). "Models of aggregate economic relationships that account for heterogeneity". In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics, vol. 6A. North-Holland, Amsterdam (Chapter 68 in this volume).

Blundell, R.W., Bond, S.R., Devereux, M.P., Schiantarelli, F. (1992). "Investment and Tobin's Q: Evidence from company panel data". Journal of Econometrics 51, 233–257.

Blundell, R.W., Bond, S.R., Meghir, C. (1996). "Econometric models of company investment". In: Matyas, L., Sevestre, P. (Eds.), The Econometrics of Panel Data. Kluwer Academic Publishers, Dordrecht.

Blundell, R.W., MaCurdy, T.E., Meghir, C. (2007). "Labor supply models: Unobserved heterogeneity, nonparticipation and dynamics". In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics, vol. 6A. North-Holland, Amsterdam (Chapter 69 in this volume).

Boal, W.M., Pencavel, J. (1994). "The effect of labor unions on employment, wages, and days of operation: Coal mining in West Virginia". Quarterly Journal of Economics 109, 267–298.

Bond, S.R., Cummins, J.G. (2000). "The stock market and investment in the new economy: Some tangible facts and intangible fictions". Brookings Papers on Economic Activity 2000 (1), 61–124.

Bond, S.R., Cummins, J.G. (2001). "Noisy share prices and the Q model of investment". Working Paper No. W01/22. The Institute for Fiscal Studies, London.

Bond, S.R., Meghir, C. (1994). "Dynamic investment models and the firm's financial policy". Review of Economic Studies 61, 197–222.

Bond, S.R., Elston, J., Mairesse, J., Mulkay, B. (2003). "Financial factors and investment in Belgium, France, Germany and the UK: A comparison using company panel data". Review of Economics and Statistics 85 (1), 153–165.

Bond, S.R., Harhoff, D., Van Reenen, J. (2003). "Investment, R&D and financing constraints in Britain and Germany". Centre for Economic Performance Discussion Paper No. 595. Annales d'Economie et de Statistique. In press.

Booth, A. (1995). The Economics of the Trade Union. Cambridge University Press, Cambridge.

Bourguignon, F., Chiappori, P.A. (1992). "Collective models of household behaviour: An introduction". European Economic Review 36, 355–365.

Brainard, W., Tobin, J. (1968). "Pitfalls in financial model building". American Economic Review 58 (2), 99–122.

Bresnahan, T.F., Brynjolfsson, E., Hitt, L.M. (2002). "Information technology, workplace organisation and the demand for skilled labor: Firm-level evidence". Quarterly Journal of Economics 117, 339–376.

Bresson, G., Kramarz, F., Sevestre, P. (1992). "Heterogeneous labour and the dynamics of aggregate labour demand: Some estimations using panel data". Empirical Economics 17, 153–168.

Brown, R.S., Christensen, L. (1981). "Estimating elasticities of substitution in a model of partial static equilibrium: An application to US agriculture 1947–1974". In: Field, C., Berndt, E. (Eds.), Modelling and Measuring Natural Resource Substitution. MIT Press, Cambridge.

Burgess, S., Dolado, J. (1989). "Intertemporal rules with variable speeds of adjustment: An application to UK manufacturing employment". Economic Journal 99, 347–365.

Caballero, R.J. (1991). "On the sign of the investment-uncertainty relationship". American Economic Review 81, 279–288.

Caballero, R.J. (1992). "A fallacy of composition". American Economic Review 82, 1279–1292.

Caballero, R.J. (1994). "Small sample bias and adjustment costs". Review of Economics and Statistics 76 (1), 52–58.

Caballero, R.J. (1999). "Aggregate investment". In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1B. North-Holland, Amsterdam.

Caballero, R.J., Engel, E. (1993). "Microeconomic adjustment hazards and aggregate dynamics". Quarterly Journal of Economics 108, 313–358.

Caballero, R.J., Engel, E. (1999). "Explaining investment dynamics in US manufacturing: A generalized, (S, s) approach". Econometrica 67, 783–826.

Caballero, R.J., Leahy, J.V. (1996). "Fixed costs: The demise of marginal q". National Bureau of Economic Research Working Paper No. 5508.

Caballero, R.J., Engel, E., Haltiwanger, J. (1995). "Plant-level adjustment and aggregate investment dynamics". Brookings Papers on Economic Activity 1995 (2), 1–54.
Caballero, R.J., Engel, E., Haltiwanger, J. (1997). "Aggregate employment dynamics: Building from microeconomic evidence". American Economic Review 87, 115–137.
Calomiris, C.W., Hubbard, R.G. (1995). "Internal finance and firm-level investment: Evidence from the undistributed profits tax of 1936–1937". Journal of Business 68, 443–482.
Campbell, J.R., Fisher, J.D.M. (2000). "Aggregate employment fluctuations with microeconomic asymmetries". American Economic Review 90, 1323–1325.
Campbell, J.Y., Kyle, A.S. (1993). "Smart money, noise trading and stock-price behaviour". Review of Economic Studies 60, 1–34.
Card, D. (1986). "Efficient contracts with costly adjustment". American Economic Review 76, 1045–1071.
Card, D., Krueger, A. (1995). Myth and Measurement. Princeton University Press, Princeton.
Caroli, E., Van Reenen, J. (2001). "Skill-biased organisational change? Evidence from a panel of British and French establishments". Quarterly Journal of Economics 116, 1449–1492.
Chambers, R. (1988). Applied Production Analysis. Cambridge University Press, Cambridge.
Chennells, L., Van Reenen, J. (1997). "Technical change and earnings in British establishments". Economica 64, 587–604.
Chesher, A. (1991). "The effect of a measurement error". Biometrika 78, 451–462.
Chirinko, R.S. (1993a). "Business fixed investment spending: Modelling strategies, empirical results and policy implications". Journal of Economic Literature 31, 1875–1911.
Chirinko, R.S. (1993b). "Multiple capital goods, Q and investment spending". Journal of Economic Dynamics and Control 17, 907–928.
Chirinko, R.S. (1997). "Finance constraints, liquidity, and investment spending: Theoretical restrictions and international evidence". Journal of the Japanese and International Economies 11, 185–207.
Chirinko, R.S., Fazzari, S.M., Meyer, A.P. (1999). "How responsive is business capital formation to its user cost? An exploration with micro data". Journal of Public Economics 74, 53–80.
Christensen, L.R., Jorgenson, D.W., Lau, L.J. (1971). "Conjugate duality and the transcendental logarithmic production function". Econometrica 39, 255–256.
Christensen, L.R., Jorgenson, D.W., Lau, L.J. (1973). "Transcendental logarithmic production frontiers". Review of Economics and Statistics 55, 28–45.
Christofides, L., Oswald, A. (1991). "Efficient and inefficient employment outcomes: A study based on Canadian contract data". Research in Labor Economics 12, 173–190.
Cobb, C., Douglas, P.H. (1928). "A theory of production". American Economic Review 18, 139–165.
Cohen, W., Levin, R. (1989). "Empirical studies of innovation and market structure". In: Schmalensee, R., Willig, R. (Eds.), The Handbook of Industrial Organisation, vol. 1. North-Holland, Amsterdam.
Cooper, R.W., Ejarque, J. (2003). "Financial frictions and investment: Requiem in Q". Review of Economic Dynamics 6, 710–728.
Cooper, R.W., Haltiwanger, J.C. (2006). "On the nature of capital adjustment costs". Review of Economic Studies 73 (3), 611–633.
Cooper, R.W., Willis, J.L. (2004). "A comment on the economics of labor adjustment: Mind the gap". American Economic Review 94 (4), 1223–1237.
Cooper, R.W., Haltiwanger, J.C., Power, L. (1999). "Machine replacement and the business cycle: Lumps and bumps". American Economic Review 84, 921–946.
Craig, B., Pencavel, J. (1994). "The empirical performance of orthodox models of the firm: Conventional firms and worker co-operatives". Journal of Political Economy 102, 718–744.
Cummins, J.G., Hassett, K.A., Hubbard, R.G. (1994). "A reconsideration of investment behaviour using tax reforms as natural experiments". Brookings Papers on Economic Activity 1994 (2), 1–59.
Cummins, J.G., Hassett, K.A., Hubbard, R.G. (1996). "Tax reforms and investment: A cross-country comparison". Journal of Public Economics 62, 237–273.
Cummins, J.G., Hassett, K.A., Oliner, S.D. (2006). "Investment behavior, observable expectations, and internal funds". American Economic Review 96 (3), 796–810.

Dagenais, M., Mohnen, P., Therrien, P. (1997). "Do Canadian firms respond to fiscal incentives to research and development?". GREQAM Working Paper 97b05.
Davidson, J.E.H., Hendry, D.F., Srba, F., Yeo, S. (1978). "Econometric modelling of the aggregate time-series relationships between consumers' expenditure and income in the United Kingdom". Economic Journal 88, 661–692.
Davis, S., Haltiwanger, J. (1992). "Gross job creation, gross job destruction and employment reallocation". Quarterly Journal of Economics 107, 819–863.
Devereux, M.P. (1989). "Tax asymmetries, the cost of capital and investment: Some evidence from UK panel data". Economic Journal 99 (Suppl.), 103–112.
Devereux, M.P., Schiantarelli, F. (1990). "Investment, financial factors and cash flow: Evidence from UK panel data". In: Hubbard, R.G. (Ed.), Asymmetric Information, Corporate Finance and Investment. University of Chicago Press.
Devereux, M.P., Keen, M.J., Schiantarelli, F. (1994). "Corporation tax asymmetries and investment: Evidence from UK panel data". Journal of Public Economics 53, 395–418.
DiNardo, J., Pischke, J.S. (1997). "The returns to computer use revisited: Have pencils changed the wage structure too?". Quarterly Journal of Economics 112, 291–303.
Dixit, A.K., Pindyck, R.S. (1994). Investment under Uncertainty. Princeton University Press.
Doms, M., Dunne, T. (1998). "Capital adjustment patterns in manufacturing plants". Review of Economic Dynamics 1 (2), 409–429.
Doms, M., Dunne, T., Troske, K. (1997). "Workers, wages and technology". Quarterly Journal of Economics 112, 253–289.
Duguet, E., Greenan, N. (1997). "Skill biased technological change: An econometric study at the firm level". CREST Working Paper.
Dunne, T., Haltiwanger, J., Troske, K. (1997). "Technology and jobs: Secular changes and cyclical dynamics". Carnegie–Rochester Conference Series on Public Policy 46, 107–178.
Eberly, J.C. (1997). "International evidence on investment and fundamentals". European Economic Review 41, 1055–1078.
Edwards, J.S.S., Kay, J.A., Mayer, C.P. (1987). The Economic Analysis of Accounting Profitability. Oxford University Press.
Eisner, R. (1977). "Cross section and time series estimates of investment functions". Annales de l'Insee 30/31, 99–129.
Elston, J. (1993). "Firm ownership structure and investment: Theory and evidence from German manufacturing". WZB Discussion Paper no. FS IV 93-28. Berlin.
Engle, R.F., Granger, C.W.J. (1987). "Cointegration and error correction: Representation, estimation, and testing". Econometrica 55, 251–276.
Erickson, T., Whited, T.M. (2000). "Measurement error and the relationship between investment and q". Journal of Political Economy 108, 1027–1057.
Favero, C.A., Pesaran, M.H. (1994). "Oil investment in the North Sea". Economic Modelling 11, 308–329.
Fazzari, S.M., Hubbard, R.G., Petersen, B.C. (1988). "Financing constraints and corporate investment". Brookings Papers on Economic Activity 1988 (1), 141–195.
Fazzari, S.M., Hubbard, R.G., Petersen, B.C. (2000). "Investment-cash flow sensitivities are useful: A comment on Kaplan and Zingales". Quarterly Journal of Economics 115 (2), 695–705.
Galeotti, M., Schiantarelli, F. (1991). "Generalised Q models for investment". Review of Economics and Statistics 73, 383–392.
Galeotti, M., Schiantarelli, F. (1994). "Stock market volatility and investment: Do only fundamentals matter?". Economica 61, 147–165.
Gera, S., Gu, W., Lin, Z. (2001). "Technology and the demand for skills: An industry-level analysis". Canadian Journal of Economics 34, 132–148.
Gilchrist, S. (1991). "An empirical analysis of corporate investment and financing hierarchies using firm-level panel data". Mimeo. Board of Governors of the Federal Reserve System.
Gilchrist, S., Himmelberg, C.P. (1995). "Evidence on the role of cash flow for investment". Journal of Monetary Economics 36, 541–572.

Gomes, J.F. (2001). "Financing investment". American Economic Review 91, 1263–1285.
Goux, N., Maurin, E. (2000). "The decline in demand for unskilled labour: An empirical analysis method and its application to France". Review of Economics and Statistics 82 (4), 596–607.
Griliches, Z. (1969). "Capital-skill complementarity". Review of Economics and Statistics 51, 465–468.
Griliches, Z. (1984). R&D, Patents and Productivity. Chicago University Press, Chicago.
Griliches, Z. (1998). R&D and Productivity: The Econometric Evidence. Chicago University Press, Chicago.
Griliches, Z., Mairesse, J. (1998). "Production functions: The search for identification". In: Strøm, S. (Ed.), Econometrics and Economic Theory in the 20th Century. Cambridge University Press, Cambridge.
Griliches, Z., Hall, B.H., Pakes, A. (1991). "R&D, patents and market value revisited: Is there a technological opportunity factor?". Economics of Innovation and New Technology 1, 183–201.
Guiso, L., Parigi, G. (1999). "Investment and demand uncertainty". Quarterly Journal of Economics 114 (1), 185–227.
Hall, B.H. (1992). "Investment and research and development at the firm level: Does the source of financing matter?". Working Paper No. 92-194. Department of Economics. University of California, Berkeley.
Hall, B.H. (1993). "R&D tax policy during the 1980s: Success of failure?". Tax Policy and the Economy 7, 1–36.
Hall, B.H., Mairesse, J., Branstetter, L., Crepon, B. (1999). "Does cash flow cause investment and R&D: An exploration using panel data for French, Japanese and United States firms in the scientific sector". In: Audretsch, D., Thurik, A.R. (Eds.), Innovation, Industry Evolution and Employment. Cambridge University Press, Cambridge.
Hall, R.E., Jorgenson, D.W. (1967). "Tax policy and investment behaviour". American Economic Review 57, 391–414.
Hamermesh, D. (1989). "Labor demand and the structure of adjustment costs". American Economic Review 79, 674–689.
Hamermesh, D. (1993). Labor Demand. Princeton Academic Press.
Hamermesh, D., Pfann, G. (1996). "Adjustment costs in factor demand". Journal of Economic Literature 34, 1264–1292.
Hansson, P. (1997). "Trade, technology and changes in employment of skilled labor in Swedish manufacturing". In: Fagerberg, J., Hansson, P., Lundberg, L., Melchior, A. (Eds.), Technology and International Trade. Edward Elgar, Cheltenham.
Harhoff, D. (1998). "Are there financing constraints for innovation and investment in German manufacturing firms?". Annales d'Économie et de Statistique 49/50, 421–456.
Haskel, J.E., Heden, Y. (1999). "Computers and the demand for skilled labour: Industry and establishment panel evidence for the UK". Economic Journal 109, 68–79.
Hassett, K.A., Hubbard, R.G. (1996). "Tax policy and investment". National Bureau of Economic Research Working Paper No. 5683.
Hayashi, F. (1982). "Tobin's average q and marginal q: A neoclassical interpretation". Econometrica 50, 213–224.
Hayashi, F. (1985a). "Corporate finance side of the Q theory of investment". Journal of Public Economics 27, 261–280.
Hayashi, F. (1985b). "Tests of liquidity constraints: A critical survey". In: Bewley, T. (Ed.), Advances in Econometrics, vol. II. Cambridge University Press.
Hayashi, F., Inoue, T. (1991). "The relation between firm growth and q with multiple capital goods: Theory and evidence from Japanese panel data". Econometrica 59 (3), 731–754.
Heckman, J.J. (1979). "Sample selection bias as a specification error". Econometrica 47, 153–161.
Heckman, J.J., Sedlacek, G. (1985). "Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market". Journal of Political Economy 93, 1077–1125.
Hendry, D.F. (1995). Dynamic Econometrics. Oxford University Press.
Himmelberg, C., Peterson, B. (1994). "R&D and internal finance: A panel data study of small firms in high tech industries". Review of Economics and Statistics 76 (1), 38–51.
Hines, J. (1994). "No place like home: Tax incentives and the location of R&D by American multinationals". Tax Policy and the Economy 8, 65–104.

Holtz-Eakin, D., Newey, W., Rosen, H.S. (1988). "Estimating vector autoregressions with panel data". Econometrica 56, 1371–1395.
Hoshi, T., Kashyap, A.K., Scharfstein, D. (1991). "Corporate structure, liquidity and investment: Evidence from Japanese industrial groups". Quarterly Journal of Economics CVI, 33–60.
Hsiao, C. (1986). Analysis of Panel Data. Cambridge University Press.
Hubbard, R.G. (1998). "Capital-market imperfections and investment". Journal of Economic Literature 36, 193–225.
Hubbard, R.G., Kashyap, A.K., Whited, T.M. (1995). "Internal finance and firm investment". Journal of Money, Credit and Banking 27, 683–701.
Jaramillo, F., Schiantarelli, F., Sembenelli, A. (1993). "Are adjustment costs for labour asymmetric? An econometric test on panel data for Italy". Review of Economics and Statistics 75, 640–648.
Jensen, M. (1986). "Agency costs of free cash flow, corporate finance, and takeovers". American Economic Review 76, 323–329.
Jorgenson, D.W. (1963). "Capital theory and investment behaviour". American Economic Review 53, 247–259.
Jorgenson, D.W. (1971). "Econometric studies of investment behaviour". Journal of Economic Literature 9, 1111–1147.
Jorgenson, D.W., Landau, R. (1993). Tax Reform and the Cost of Capital. Brookings Institution, Washington, DC.
Kaplan, S.N., Zingales, L. (1997). "Do investment-cash flow sensitivities provide useful measures of financing constraints?". Quarterly Journal of Economics 112 (1), 169–216.
Kaplan, S.N., Zingales, L. (2000). "Investment-cash flow sensitivities are not valid measures of financing constraints". Quarterly Journal of Economics 115 (2), 707–712.
King, M.A. (1974). "Taxation and the cost of capital". Review of Economic Studies 41, 21–35.
King, M.A., Fullerton, D. (1984). The Taxation of Income from Capital. University of Chicago Press.
Klette, T.J., Moen, J. (1999). "From growth theory to technology policy: Coordination problems in theory and practice". Nordic Journal of Political Economy 25, 53–74.
Krueger, A. (1993). "How have computers changed the wage structure: Evidence from micro data, 1984–1989". Quarterly Journal of Economics 108, 33–60.
Lach, S., Rob, R. (1996). "R&D, investment and industry dynamics". Journal of Economics and Management Strategy 5 (2), 217–249.
Lach, S., Schankerman, M. (1989). "The dynamics of R&D investment in the scientific sector". Journal of Political Economy 97, 880–904.
Lamont, O.A. (1997). "Cash flow and investment: Evidence from internal capital markets". Journal of Finance 52, 83–109.
Leahy, J.V., Whited, T.M. (1996). "The effect of uncertainty on investment: Some stylised facts". Journal of Money, Credit and Banking 28, 64–83.
Leontief, W. (1946). "The pure theory of the guaranteed annual wage contract". Journal of Political Economy 54, 76–79.
Lucas, R.E. (1967). "Adjustment costs and the theory of supply". Journal of Political Economy 75, 321–334.
Lucas, R.E. (1976). "Econometric policy evaluation: A critique". In: Brunner, K., Meltzer, A. (Eds.), The Phillips Curve and Labour Markets. In: Carnegie–Rochester Conference Series, vol. 1, pp. 19–46.
Machin, S. (1996). "Changes in the relative demand for skills in the UK labour market". In: Booth, A., Snower, D. (Eds.), Acquiring Skills. Cambridge University Press, Cambridge.
Machin, S., Van Reenen, J. (1998). "Technology and changes in the skill structure: Evidence from seven OECD countries". Quarterly Journal of Economics 113, 1215–1244.
Machin, S., Manning, A., Meghir, C. (1993). "Dynamic models of employment based on firm level panel data". In: Van Ours, J., Pfann, G., Ridder, G. (Eds.), Labor Demand and Equilibrium Wage Formation. Elsevier Science Publishers, Amsterdam.
MaCurdy, T., Pencavel, J. (1986). "Testing between competing models of wage and employment determination". Journal of Political Economy 94, S1–S39.

Mairesse, J., Dormont, B. (1985). "Labor and investment demand at the firm level: A comparison of French, German and US manufacturing, 1970–1979". European Economic Review 28, 201–231.
Manning, A. (1987). "An integration of trade-union models in a sequential bargaining framework". Economic Journal 97, 121–139.
Meghir, C., Ryan, A., Van Reenen, J. (1996). "Job creation, technological innovation and adjustment costs". Annales d'Économie et de Statistique 41/42, 255–274.
Menezes-Filho, N., Ulph, D., Van Reenen, J. (1998). "R&D investment and union bargaining: Evidence from British companies and establishments". Industrial and Labor Relations Review 52 (1), 45–63.
Miller, M.H., Modigliani, F. (1961). "Dividend policy, growth and the valuation of shares". Journal of Business 44, 411–433.
Modigliani, F., Miller, M.H. (1958). "The cost of capital, corporation finance, and the theory of investment". American Economic Review 48, 261–297.
Mulkay, B., Hall, B.H., Mairesse, J. (2001). "Investment and R&D in France and the United States". In: Heinz, H., Strauch, R. (Eds.), Investing Today for the World of Tomorrow. Springer-Verlag.
Myers, S.C. (1984). "The capital structure puzzle". Journal of Finance 39, 575–592.
Nadiri, M., Rosen, S. (1969). "Interrelated factor demand functions". American Economic Review 59, 547–571.
Nelson, C.R., Startz, R. (1990a). "Some further results on the exact small sample properties of the instrumental variable estimator". Econometrica 58, 967–976.
Nelson, C.R., Startz, R. (1990b). "The distribution of the instrumental variable estimator and its t-ratio when the instrument is a poor one". Journal of Business and Economic Statistics 63, 5125–5140.
Nickell, S.J. (1978). The Investment Decisions of Firms. Cambridge University Press, Cambridge.
Nickell, S.J. (1981). "Biases in dynamic models with fixed effects". Econometrica 49, 1417–1426.
Nickell, S.J. (1985). "Error correction, partial adjustment and all that: An expository note". Oxford Bulletin of Economics and Statistics 47 (2), 119–129.
Nickell, S.J. (1986). "Dynamic models of labor demand". In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, vol. 1. North-Holland, Amsterdam.
Nickell, S.J., Nicolitsas, D. (1996). "Does innovation encourage investment in fixed capital?" Centre for Economic Performance Discussion Paper No. 309. London School of Economics.
Nickell, S.J., Wadhwani, S. (1991). "Employment determination in British industry: Investigations using micro-data". Review of Economic Studies 58, 955–970.
Olley, G.S., Pakes, A. (1996). "The dynamics of productivity in the telecommunications equipment industry". Econometrica 64 (6), 1263–1297.
Pakes, A. (1985). "On patents, R&D and the stock market rate of return". Journal of Political Economy 93, 390–409.
Pencavel, J. (1991). Labor Markets Under Trade Unionism. Oxford University Press, Oxford.
Pesaran, M.H. (1990). "An econometric analysis of exploration and extraction of oil in the UK continental shelf". Economic Journal 100, 367–390.
Pesaran, M.H., Smith, R. (1995). "Estimating long-run relationships from dynamic heterogeneous panels". Journal of Econometrics 68, 79–113.
Pfann, G., Palm, F. (1993). "Asymmetric adjustment costs in non-linear labour demand models of the Netherlands and UK manufacturing sectors". Review of Economic Studies 60, 397–412.
Pfann, G., Verspagen, B. (1989). "The structure of adjustment costs for labour in the Dutch manufacturing sector". Economics Letters 29, 365–371.
Romer, P.M. (1986). "Increasing returns and long run growth". Journal of Political Economy 94 (5), 1002–1037.
Rota, P. (1998). "Dynamic labour demand with fixed costs of adjustment". Mimeo. University College London.
Salinger, M.A., Summers, L.H. (1983). "Tax reform and corporate investment: A microeconometric simulation study". In: Feldstein, M. (Ed.), Behavioural Simulation Methods in Tax Policy Analysis. University of Chicago Press.

Sargan, J.D. (1964). "Wages and prices in the UK: A study in econometric methodology". In: Hart, P.E., Mills, G., Whitaker, J.K. (Eds.), Econometric Analysis for National Economic Planning. Butterworths, London.

Schaller, H. (1993). "Asymmetric information, liquidity constraints and Canadian investment". Canadian Journal of Economics 26, 552–574.

Schiantarelli, F. (1996). "Financial constraints and investment: A critical survey of the international evidence". Oxford Review of Economic Policy 12 (2), 70–89.

Schiantarelli, F., Georgoutsos, D. (1990). "Imperfect competition, Tobin's q and investment: Evidence from aggregate UK data". European Economic Review 34, 1061–1078.

Schiantarelli, F., Sembenelli, A. (1993). "Estimation of Euler equations for employment when adjustment costs are asymmetric: Evidence from panel data for UK companies". In: Van Ours, J., Pfann, G., Ridder, G. (Eds.), Labor Demand and Equilibrium Wage Formation. Elsevier Science Publishers, Amsterdam.

Shapiro, M.D. (1986). "The dynamic demand for capital and labor". Quarterly Journal of Economics CI, 513–542.

Shephard, R. (1953). Cost and Production Functions. Princeton University Press.

Staiger, D., Stock, J.H. (1997). "Instrumental variables regression with weak instruments". Econometrica 65, 557–586.

Stiglitz, J.E., Weiss, A. (1981). "Credit rationing in markets with imperfect information". American Economic Review 71, 393–410.

Summers, L.H. (1981). "Taxation and corporate investment – a Q-theory approach". Brookings Papers on Economic Activity 1981 (1), 67–140.

Summers, L.H. (1986). "Does the stock market rationally reflect fundamental values?". Journal of Finance 41, 591–600.

Sutton, J. (1998). Technology and Market Structure. MIT Press, Cambridge.

Tobin, J. (1969). "A general equilibrium approach to monetary theory". Journal of Money, Credit and Banking 1, 15–29.

Toivanen, O., Stoneman, P. (1998). "Dynamics of R&D and investment: UK evidence". Economic Letters 58, 119–126.

Uzawa, H. (1962). "Production functions with constant elasticities of substitution". Review of Economic Studies 29, 291–299.

Weiss, R. (1977). "Elasticities of substitution across capital and occupations in US manufacturing". Journal of the American Statistical Association 72, 764–771.

Whited, T.M. (1992). "Debt, liquidity constraints and corporate investment: Evidence from panel data". Journal of Finance 47, 1425–1460.

Zeldes, S.P. (1989). "Consumption and liquidity constraints: An empirical investigation". Journal of Political Economy 97, 305–346.

Chapter 66

THE MEASUREMENT OF PRODUCTIVITY FOR NATIONS*

W. ERWIN DIEWERT

Department of Economics, University of British Columbia, Vancouver, British Columbia V6T 1W5, Canada
e-mail: diewert@econ.ubc.ca

ALICE O. NAKAMURA

School of Business, University of Alberta, Edmonton, Alberta T6G 2R6, Canada
e-mail: alice.nakamura@ualberta.ca

Contents

* This research was funded in part by the grants from the Social Sciences and Humanities Research Council of Canada (SSHRC). Thanks for discussions and for comments on various earlier versions of this paper are due to James Heckman, Rosa Matzkin, Amil Petrin, Arnold Zellner, Richard Blundell and other participants in the University of Chicago *Handbook of Econometrics*, Volume 6 Conference; to Bo Honoré and other participants in a seminar held at the Princeton University Economics Department; and to Andy Baldwin, John Baldwin, Bert Balk, Susanto Basu, Jeff Bernstein, Pierre Duguay, Bob Fay, Kevin Fox, Mel Fuss, T. Peter Hill, Robert H. Hill, Ulrich Kohli, David Laidler, Denis Lawrence, Frank Lee, Richard Lipsey, J.P. Maynard, Takanobu Nakajima, Emi Nakamura, Leonard Nakamura, Masao Nakamura, Koji Nomura, Marc Prud'homme, Someshwar Rao, Paul Schreyer, Jón Steinsson, Jianmin Tang, Jack Triplett, Tom Wilson and Kam Yu as well as other participants in a series of workshops held at Statistics Canada and in a Union College workshop. All errors are the sole responsibility of the authors.

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06066-7

Abstract

This chapter covers the theory and methods for productivity measurement for nations. Labor, multifactor and total factor productivity measures are defined and are related to each other and to gross domestic product (GDP) per capita. Their growth over time and relative counterparts are defined as well.

Different conceptual meanings have been proposed for a total factor productivity growth (TFPG) index. These are easiest to understand for the case in which the index number problem is absent: a production process that involves one input and one output (a 1–1 process). It is easily seen that four common concepts of TFPG all lead to the same result in the 1–1 case. Moving on to a general N input, M output production scenario, we demonstrate that a Paasche, Laspeyres or Fisher index number formula provides a measure for all four of the concepts of TFPG introduced for the 1–1 case. This is an advantage of the Paasche–Laspeyres–Fisher family of formulas.

When multiple inputs or outputs are involved, there is the problem of choosing among alternative functional forms. The axiomatic and economic approaches to index formula choice are reviewed.

In addition, we briefly cover the Divisia index number approach and growth accounting, including the KLEMS (capital, labor, energy, materials and services) approach. The gross output measures of the KLEMS approach are contrasted with value added output

measures such as GDP. Also, an alternative family of revenue function based productivity growth indexes proposed by Diewert, Kohli and Morrison (DKM) is outlined. The DKM approach facilitates the decomposition of productivity growth into economically meaningful components. This approach is useful, for example, for examining the effects of changes in the terms of trade on productivity growth.

Keywords

total factor productivity growth, labor productivity, living standards, exact index numbers, capital deepening, real income growth, gross versus net output, growth accounting, KLEMS, terms of trade, aggregation of capital, embodiment of technical progress, depreciation, deterioration, obsolescence, index number theory

JEL classification: O4, O4.7, C43, C82, D24, D31, F14, I3

1. Introduction

> "The two main sources of economic growth in output are increases in the factors of production (the labor and capital devoted to production) and efficiency or productivity gains that enable an economy to produce more for the same amount of inputs."
>
> [Baldwin, Harchaoui, Hosein and Maynard (2001), "Productivity: Concepts and Trends", Statistics Canada]

> "Productivity is commonly defined as a ratio of a volume measure of output to a volume measure of input use. While there is no disagreement on this general notion, a look at the productivity literature and its various applications reveals very quickly that there is neither a unique purpose for nor a single measure of productivity."
>
> [Paul Schreyer (2001), OECD Productivity Manual]

Productivity for nations is like love. Much is said about the benefits of having more of it, but consensus is elusive on what "it" really is. As Schreyer (2001) writes, "productivity is commonly defined as a ratio of a volume measure of output to a volume measure of input use." But how can the output and input volumes be defined and measured for a nation? This paper deals with the methods used for measuring aggregate productivity, by which we mean the productivity of unique entities such as nations or entire industries.

The best of all times for reviewing a subject area is when the reported findings are impacting important decision processes, so the research matters; when there is a large volume of recent research to be digested and integrated with previous findings; when important data developments have taken place or are in progress; and when there is informed and truly interactive debate on how best to proceed in areas where researchers disagree on the appropriate directions. This is the current state of affairs for the subject of this paper: the measurement of productivity for nations.

Multiple types of productivity measures are produced for nations. Official statistics agencies in countries, including the United States and Canada, produce three sorts of labor productivity measures. In this paper we refer to these using the designations of per worker labor productivity (LP), per hour labor productivity (HLP), and per weighted hour labor productivity (WHLP).

Many official statistics agencies also produce a multifactor productivity measure (MFP) that takes account of machinery and equipment and other capital inputs as well as labor, and sometimes energy and materials inputs as well. Even though it is probably not possible to measure *all* inputs at a national level, economists define and consider and estimate approximations to total factor productivity (TFP) measures.

In the rest of this paper we focus mostly on the TFP and TFPG measures, where we use TFPG to denote both TFP growth and relative TFP. (Note that others sometimes make this same distinction by using TFP for both the growth and relative total factor

productivity measures with the qualifier of "levels" in referring to what we denote as simply TFP.)

National productivity estimates are of special importance because they are an input into many aspects of policy making.[1] Although useful analogies can be drawn and there are methodological commonalities, the measurement of productivity for nations is a fundamentally different undertaking from the sorts of productivity measurement dealt with by engineers for specific machines and production lines, and by accountants and business analysts and economists working with micro level data for individual production units. At this level of aggregation, the data available are limited to fairly short time series, putting bounds on the scope for econometric estimation. Also feedback effects among the measured inputs and outputs cannot be ruled out *a priori*. Index number methods (including growth accounting) are the mainstay methodology.

Estimates of relative productivity or productivity growth do not, by themselves, provide causal insights. However, many aspects of federal government and other economic planning are affected by reported productivity measures. Also, causal research on productivity depends as well on having measures of productivity.[2]

Many economists seem not to look on index number theory and applied research as belonging within the discipline of economics. And yet, there is scarcely an empirical paper published in economics that does not utilize price or volume, if not productivity, index numbers. Certainly index numbers are ubiquitous in empirical macroeconomics.

Also, economic theory and empirical findings provide the basis for a wide array of choices made in defining and evaluating national price, volume and productivity indexes. For example, the exact index number method for choosing among alternative index number formulas involves showing that specific ones can be derived from optimizing models for firms or households where these models include production, revenue,

[1] A sense of the range of relevant public policy issues can be acquired from studies including Aschauer (1989), Atrostic and Nguyen (2006), Baily (1981), Baldwin, Jarmin and Tang (2004), Balk (1998, 2003), Basu and Fernald (1997), Basu et al. (2004), Berndt and Wood (1975), Black and Lynch (1996), Boskin (1997), Bresnahan and Gordon (1997), Denison (1979), Diewert (1993a, 1995, 1998a, 1998b, 2001a, 2001c, 2002a, 2005e, 2005f, 2006c, 2007b), Diewert and Fox (1999), Diewert and Lawrence (2005), Diewert, Lawrence and Fox (2006), Duguay (1994, 2006), Ellerman, Stoker and Berndt (2001), Feenstra and Hanson (2005), Fortin (1996), Griliches (1997), Ho, Rao and Tang (2004, 2007), Hulten (1986, 2001), Jog and Tang (2001), Jorgenson (2001, 2004), Jorgenson, Ho and Stiroh (2005), Jorgenson and Landefeld (2006), Jorgenson and Lee (2001), Jorgenson and Motohashi (2005), Jorgenson and Nomura (2005), Jorgenson and Yun (1986, 1990, 1991), Kuroda and Nomura (2003), Lee and Tang (2001a, 2001b), Lipsey and Carlaw (2004), Lipsey, Carlaw and Bekar (2006), Nakamura and Lipsey (2006), Maddison (1987), Mankiw (2001), Morrison (1992), Muellbauer (1986), Nadiri (1980), Nakamura and Diewert (2000), Nordhaus (1982), Power (1998), Prescott (1998), Smith (2005), Stiroh (2002) Tang and Wang (2004, 2005), Triplett and Bosworth (2004), van Ark, Inklaar and McGuckin (2003), and Wolff (1996).

[2] For gaining a causal understanding of the determinants of national productivity, data at lower levels of aggregation are of obvious value, as are suitable econometric methods for analyzing panel and other sorts of micro data files. See, for example, Bartelsman and Doms (2000), Foster, Haltiwanger and Krizan (2001), Levinsohn and Petrin (1999), Olley and Pakes (1996), and Pavcnik (2002).

cost, expenditure, transformation, or other aggregators with specific functional forms such as the translog and the generalized quadratic with properties that have been explored by economists. We feel that index number theory and practice should (once more) be a core subject within economics.

The traditional index number measures of TFPG are defined as ratios of output and input volume indexes. As is appropriate, statistics agencies collect mostly value and price, rather than volume and price, information, and then create the needed volume data by deflating value data using price indexes. We show the relationships among price, volume and productivity indexes, and how productivity indexes relate to real revenue/cost ratios.

Several different conceptual meanings have been proposed for a TFPG index. The alternative concepts are easiest to understand for a one period production process that uses a single input factor to make a single output product (a 1–1 process). In Section 2 we show that four common concepts of TFPG all lead to the same measure in the 1–1 case. Of course, the aggregation challenges that must be confronted in the construction of national productivity measures do not arise in a 1–1 case context. To introduce these issues, we use a hypothetical two input, one output production scenario (a 2–1 process). We then move on to the general N input, M output case that is relevant for national level productivity measurement.

In the final subsection of Section 2 we introduce three different labor productivity indexes in common use, and relate these to the multifactor productivity (MFP) and total factor productivity (TFP) measures that are our main focus in this paper. The Törnqvist, and implicit Törnqvist volume and price indexes[3] and the corresponding TFPG indexes are also introduced and discussed.

In Section 3 we define Laspeyres, Paasche and Fisher measures for the general N input, M output case for the four concepts of TFPG introduced in Section 2 for the 1–1 case.

With multiple inputs and outputs, different formula choices lead to different TFPG findings. This raises the issue of choice among alternative TFPG formulas. The two main approaches to choosing among the different index number functional forms are the axiomatic (or test) approach and the economic approach.[4]

The axiomatic approach is taken up in Section 4. It was used extensively by the founding contributors to index number theory, including Fisher (1911, 1922). This approach makes use of lists of desired properties referred to as axioms or tests. They are either formalizations of common sense properties of good index numbers or generalizations of properties that hold for virtually all proposed index number formulas in the simplistic 1–1 case.

[3] Perhaps the best source for learning about or checking details of price indexes are the new international Consumer Price Index Manual [T.P. Hill (2004)] and Producer Price Index Manual [Armknecht (2004)]. The Diewert chapters in the new International CPI and PPI Manuals are Chapters 15–20 and 22–23 of the CPI Manual and Chapters 15–22 of the PPI Manual. See also Diewert (2002b).

[4] A third approach – the statistical approach – is not discussed here. See Diewert (1981a, 1987, 2002b, 2004c, 2007c) on this parallel approach to the index number formula choice problem.

The axiomatic approach to index number choice focuses on properties of the index number formula itself. In contrast, the economic approach seeks to use principles and implications of economic theory as a basis for choosing among proposed index formulas.

Exact index number theory is one important stream within the economic approach to index numbers: the stream on which we focus in Section 5. The exact approach transforms the index number choice problem into a problem of choosing the correct functional form for a behavioral aggregator function of some sort. In order to use the exact approach to derive the functional form for a TFPG index, it is first necessary to decide on the perspective for the productivity analysis. When a producer perspective is adopted, then the aggregator function for the economic approach can be the production function, or it can be the corresponding cost, profit, or other dual representation of the production process. Once the form of the aggregator has been determined, then the exact index number approach can be applied in order to determine the corresponding functional form for the TFPG index, as shown in Section 5.

When it can be established that some particular index number formula corresponds, by the "exact" index number approach, to a linearly homogeneous producer behavioral relationship that is "flexible", meaning that it provides a second order approximation to an arbitrary twice continuously differentiable linearly homogeneous function, then the index number is said to be "superlative". Diewert established that, under ordinary conditions, all of the commonly used superlative index number formulas (including the Fisher, Törnqvist, and implicit Törnqvist formulas introduced in Section 3) approximate each other to the second order when evaluated at an equal price and volume point. Diewert established as well that the two most commonly used index number formulas that are not superlative – the Laspeyres and the Paasche indexes – approximate the superlative indexes to the first order at an equal price and volume point.

The exact index number approach, together with Diewert's numerical analysis approximation results for superlative index numbers, reduces the *a priori* information requirements for choosing an index number formula to a list of general characteristics of the production scenario. So long as there is agreement on those characteristics, under ordinary conditions, any one of the commonly used superlative TFPG index number formulas should provide a reasonable estimate to the theoretical Malmquist TFPG index introduced in Section 6.

The exact and the axiomatic approaches single out some of the same index number formulas as especially desirable. The exact approach can be viewed as a methodology for exploring the meaning of the proposed measures of TFPG and also of the intuitions on which the axiomatic approach is based. This approach helps us interpret TFPG indexes in the language of neoclassical theory. That the index number formulas which have been in use since the early 1900s have interpretations in the language of modern microeconomic theory suggests that the intuitions which guided the axiomatic approach to index number theory and the axioms of microeconomic theory may have more in common than is readily apparent.

The data used in evaluating measures of productivity are discrete. Nevertheless, various properties of national productivity measures have been worked out utilizing the convergence of continuous approximations. The Divisia method reviewed in Section 8 treats time as continuous. The Divisia method has been used extensively in growth accounting studies for nations, which is the subject of Section 9. Section 9 also briefly takes up the KLEMS (capital, labor, energy, materials and services) approach, and the World KLEMS data development and analysis initiatives.

In Section 10, further consideration is also given to the choice of the measure of output incorporated into productivity analyses and we review efforts to relax the assumption of constant returns to scale.

In Section 11, an alternative family of theoretical productivity growth indexes proposed by Diewert and Morrison (1986) and Kohli (1990) is introduced.[5] This approach has special advantages for examining the components of TFP growth.

Section 12 concludes.

2. Alternative productivity measurement concepts

"**Productivity**

A ratio of output to input."

[Atkinson, Kaplan and Young (1995, p. 514)]

"While, for example, we look at the cost of power as a number of 'analysed' items such as coal, water-rate, ash removal, drivers' and stokers' wages, etc., it will probably be a long time before it dawns upon us that all this expenditure can be reduced to a horse-power-hour rate, and that such a factor, once known, may turn out to be a standing reproach. The burning of 200 tons of coal per week may mean anything or nothing, but the cost of a horse-power hour can be compared at once with standard data ... the publication of figures based on them would reveal amazing inefficiencies that under present conditions are unsuspected and unknown because no means of comparison exists."

[A. Hamilton Church (1909, p. 190)]

The basic definition of total factor productivity (TFP) is the rate of transformation of total input into total output. The output-over-input index approach to the measurement of total factor productivity has early origins.[6] In his Simon Kuznets Memorial Lecture, Griliches remarked that "the first mention of what might be called an output-over-input index that I can find appears in Copeland (1937)". However, in an endnote to the written version of the lecture Griliches (1997) writes:

[5] This approach has been used lately in a growing number of other studies such as Feenstra et al. (2005).

[6] Output over input measures are sometimes referred to as productivity levels measures.

> "Nothing is really new. Kuznets (1930) used the 'cost of capital and labor per pound of cotton yarn,' the inverse of what would later become a total factor productivity index (if the cost is computed in constant prices) ... as a '(reflection of) the economic effects of technical improvement' and a few sentences later as a measure of 'the effect of technical progress' (p. 14). More thorough research is likely to unearth even earlier references".

Indeed, the early engineering and cost accounting literature contains numerous references to unit costs used as efficiency measures (e.g., [Church (1909)]). For a one output production process, the unit cost is the reciprocal of the TFP index.

Virtually all real production processes make use of multiple inputs and most yield multiple outputs. Nevertheless, it is convenient to introduce basic concepts, terms and notation in the simplified context of a production process with a single homogeneous input factor and a single homogeneous output product. In a 1–1 context, the concepts of total factor productivity and total factor productivity growth (TFPG) are easy to think about because the measures are not complicated by choices about how different types of inputs and different types of outputs should be aggregated. By the same token, of course, the aggregation difficulties that arise when there are multiple inputs or outputs cannot be introduced in a 1–1 context because they do not arise. Thus in Subsection 2.2 we also briefly consider a two input, one output process, a 2–1 case before moving on in Subsection 2.3 to a general N input, M output setting. Labor, multifactor, and total factor productivity measures are introduced in Subsection 2.3.

2.1. The 1–1 case[7]

For each time period (or scenario), suppose we know the volume of the one input used, given by x_1^t, its unit price w_1^t, and the volume of the one output produced, given by y_1^t, and its unit price p_1^t. TFP can be defined conceptually as the rate of transformation of total input into total output. For the 1–1 case, the ratio of output produced to input used is the measure for TFP for period t:

$$\text{TFP} \equiv \left(y_1^t / x_1^t\right) \equiv a^t. \tag{2.1-1}$$

The parameter a^t that is defined as well in (2.1-1) is a conventional output–input coefficient.[8]

Total factor productivity growth, or TFPG, can be defined in several ways, four of which are considered here.[9] Our first concept of TFPG is the rate of transformation of

[7] This section and some of what follows draws on Diewert (2000).

[8] An output–input coefficient always involves just one output and one input. However, these coefficients can be defined and used in multiple input, multiple output situations too, as is done in Diewert and Nakamura (1999).

[9] Some authors also use TFP to refer to total factor productivity growth. In line with Bernstein (1999), we use TFPG rather than TFP for total factor productivity growth so as to avoid the inevitable confusion that otherwise results.

input into output for production period t versus s, where s comes before t here and elsewhere in this paper if these are time periods. This concept of TFPG, denoted here by TFPG(1), can be measured in the 1–1 case as[10]:

$$\text{TFPG(1)} \equiv \left(\frac{y_1^t}{x_1^t}\right)\Big/\left(\frac{y_1^s}{x_1^s}\right) = a^t/a^s. \tag{2.1-2}$$

Three other concepts of total factor productivity growth are also in common use:

- the ratio of the output and the input growth rates, denoted by TFPG(2);
- the rate of growth in the real revenue/cost ratio; i.e., the rate of growth in the revenue/cost ratio controlling for price change, denoted by TFPG(3); and
- the rate of growth in the margin after controlling for price change, denoted by TFPG(4).

For a 1–1 production process, the obvious measure for the second concept of TFPG is:

$$\text{TFPG(2)} \equiv \left(\frac{y_1^t}{y_1^s}\right)\Big/\left(\frac{x_1^t}{x_1^s}\right). \tag{2.1-3}$$

The third and fourth concepts of TFPG are financial in nature. Expressions for actual revenue and cost are needed to form measures for these. For the 1–1 case, total revenue and total cost are given by

$$R^t \equiv p_1^t y_1^t \quad \text{and} \quad C^t \equiv w_1^t x_1^t. \tag{2.1-4}$$

Thus, the third concept of TFPG can be measured by

$$\text{TFPG(3)} \equiv \left[\frac{R^t/R^s}{p_1^t/p_1^s}\right]\Big/\left[\frac{C^t/C^s}{w_1^t/w_1^s}\right] = \left(\frac{y_1^t}{y_1^s}\right)\Big/\left(\frac{x_1^t}{x_1^s}\right), \tag{2.1-5}$$

where

$$\left(R^t/R^s\right)\Big/\left(p^t/p^s\right) = \left(p_1^t y_1^t/p_1^s y_1^s\right)\Big/\left(p_1^t/p_1^s\right) = y_1^t/y_1^s, \quad \text{and} \tag{2.1-6}$$

$$\left(C^t/C^s\right)\Big/\left(w^t/w^s\right) = \left(w_1^t x_1^t/w_1^s x_1^s\right)\Big/\left(w_1^t/w_1^s\right) = x_1^t/x_1^s. \tag{2.1-7}$$

Business managers are usually interested in ensuring that revenues exceed costs, and this leads to an interest in margins. The period t margin, m^t, can be defined by

$$1 + m^t \equiv R^t/C^t. \tag{2.1-8}$$

Using this definition, in the 1–1 case TFPG(4) can be measured by

$$\text{TFPG(4)} \equiv \left[\left(1+m^t\right)\Big/\left(1+m^s\right)\right]\left[\left(w_1^t/w_1^s\right)\Big/\left(p_1^t/p_1^s\right)\right]. \tag{2.1-9}$$

[10] Here we refer to t and s as time periods. However, the 'period s' comparison situation could be for some other unit of production in the same time period.

If we interpret the margin as a reward for managerial or entrepreneurial input, then TFPG(4) can be interpreted as the rate of growth of input prices, broadly defined so as to include managerial and entrepreneurial input, divided by the rate of growth of output prices. Note that if the margins are zero, then TFPG(4) reduces to $(w_1^t/w_1^s)/(p_1^t/p_1^s)$.[11]

Using (2.1-8) to eliminate the margin growth rate on the right-hand side of (2.1-9), and comparing the resulting expression and those in (2.1-2), (2.1-3) and (2.1-5), it can readily be seen that the four concepts of total factor productivity growth introduced here all lead to the same pure volume measure. That is, for the 1–1 case the measures for all four of the concepts for TFPG reduce to

$$\text{TFPG} \equiv \left(\frac{y_1^t}{y_1^s}\right) \Big/ \left(\frac{x_1^t}{x_1^s}\right). \tag{2.1-10}$$

2.2. The 2–1 case

We next use a slightly more complex production process as the context for introducing key choices that must be faced in order to specify multiple input, multiple output measures of TFP and TFPG. This hypothetical 2–1 production process uses the labor hours of one man and logs as inputs and yields firewood as the output. The man buys the loads of logs, splits them with an axe, and then sells the split logs as firewood. The axe was inherited and has no resale or rental value. The man's time, in hours, is denoted by x_1^t, and the number of truckloads of logs purchased is denoted by x_2^t. The firewood output is measured in kilograms and denoted by y_1^t.

The labor productivity in each period is given by (y_1^t/x_1^t). The materials utilization productivity can also be defined as (y_1^t/x_2^t). These are the two output–input coefficient measures that can be specified for this production scenario, and their values will tend to move in opposite directions from period to period. When the man splits logs at a faster pace, unless he pays extra attention, he uses the raw resource input more wastefully. The fact that the single factor productivity measures do not necessarily move together closely (or even in the same direction) is a key reason why TFP and TFPG measures are needed instead of just labor productivity measures.

In order to measure TFP for our log splitting process, a measure for total input is needed. That is, we need a way of adding hours of labor and truckloads of logs. Different perspectives can be adopted for forming this aggregate.[12]

In the economic approach to index number theory, the goal of producer revenue or cost optimization dictates that unit revenues or costs should be used as weights in aggregating the volumes of the different inputs and outputs.

[11] One set of conditions under which the margins will be zero is perfect competition and constant returns to scale.

[12] This issue of perspective is taken up, for example, in Schultze and Mackie (2002).

In our firewood production example, if the unit cost for an hour of labor is w_1^t and the unit cost of a load of logs is w_2^t, then the input volume aggregate could be defined as the following price weighted sum:

$$w_1^t x_1^t + w_2^t x_2^t. \tag{2.2-1}$$

If the total input is measured as in (2.2-1), then total factor productivity, defined as the rate of transformation of total input into total output, can be measured as

$$\text{TFP} = y_1^t \big/ \big(w_1^t x_1^t + w_2^t x_2^t\big). \tag{2.2-2}$$

Now, suppose we want to measure TFPG. That is, suppose we want to compare the ratio of output to input in period (or scenario) t with the ratio of output to input for some earlier period (or some different production scenario) s. Should period t price weights be used in forming both the period t and period s aggregates? Or, should period s price weights be used in forming both of the aggregates? Or, should some sort of combination of the period s and t prices be used as weights? Also, are there other functional forms besides the linear one that might be preferable for combining the volumes of the different inputs? These are the sorts of issues that are faced in the theory of index numbers when it comes to choosing among alternative functional forms that have been proposed for the indexes.

2.3. Different types of measures of productivity

For nations, a general N input, M output production setting applies. In the next sections, we introduce formulas. Here, however, we first show how TFP and TFPG measures fit with other general types of productivity measures that are commonly used at a national level and with per capita gross domestic product (GDP). We do this here using words rather than mathematical expressions for the relevant component parts.

GDP per capita equals the product of GDP per hour of work, the average hours of work per worker, the employment rate, and the proportion of the population (denoted by POP) that is old enough to work and hence in the potential labor force:

$$\frac{\text{GDP}}{\text{POP}} \equiv \frac{\text{GDP}}{\left[\begin{array}{c}\text{Total}\\ \text{work}\\ \text{hours}\end{array}\right]} \times \frac{\left[\begin{array}{c}\text{Total}\\ \text{work}\\ \text{hours}\end{array}\right]}{\left[\begin{array}{c}\text{Number}\\ \text{of}\\ \text{workers}\end{array}\right]} \times \frac{\left[\begin{array}{c}\text{Number}\\ \text{of}\\ \text{workers}\end{array}\right]}{\left[\begin{array}{c}\text{Potential}\\ \text{labor}\\ \text{force}\end{array}\right]} \times \frac{\left[\begin{array}{c}\text{Potential}\\ \text{labor}\\ \text{force}\end{array}\right]}{\text{POP}}. \tag{2.3-1}$$

Variants of the above identity have been used in many published studies. For understanding the commonly used measures of productivity, it is useful to expand this expression as follows:

$$\frac{\text{GDP}}{\text{POP}} \equiv \frac{\text{GDP}}{\left[\begin{array}{c}\text{Total}\\ \text{input}\end{array}\right]} \times \frac{\left[\begin{array}{c}\text{Total}\\ \text{input}\end{array}\right]}{\left[\begin{array}{c}\text{Total}\\ \text{measured}\\ \text{input}\end{array}\right]} \times \frac{\left[\begin{array}{c}\text{Total}\\ \text{measured}\\ \text{input}\end{array}\right]}{\left[\begin{array}{c}\text{Total}\\ \text{labor}\\ \text{input}\end{array}\right]} \times \frac{\left[\begin{array}{c}\text{Total}\\ \text{labor}\\ \text{input}\end{array}\right]}{\left[\begin{array}{c}\text{Total}\\ \text{work}\\ \text{hours}\end{array}\right]} \times \frac{\left[\begin{array}{c}\text{Total}\\ \text{work}\\ \text{hours}\end{array}\right]}{\left[\begin{array}{c}\text{Number}\\ \text{of}\\ \text{workers}\end{array}\right]}$$

$$\times \frac{\left[\begin{array}{c}\text{Number}\\ \text{of}\\ \text{workers}\end{array}\right]}{\left[\begin{array}{c}\text{Potential}\\ \text{labour}\\ \text{force}\end{array}\right]} \times \frac{\left[\begin{array}{c}\text{Potential}\\ \text{labour}\\ \text{force}\end{array}\right]}{\text{POP}}$$

$$= (\text{A}) \times (\text{B}) \times (\text{C}) \times (\text{D}) \times (\text{E}) \times (\text{F}) \times (\text{G}). \tag{2.3-2}$$

For expositional convenience, we denote the terms on the right-hand side by A–G, respectively.

All of the productivity measures we consider have as their numerator some measure of total output. We follow common practice here in using GDP as the measure of national output.[13] On a conceptual level, productivity is just output over input – that is, it is the rate of conversion of input into output. These various productivity measures differ in terms of the categories of included input.[14]

Productivity measures in common use and our designations for these are total factor productivity (TFP), multi factor productivity (MFP), labor productivity with wage weighted hours of work used as the measure of labour input (WHLP), labor productivity with hours of work used as the measure of labor input (what we denote here as HLP), and labor productivity with the number of workers used as the measure of labor input (LP).

To be meaningfully interpreted, productivity measures must usually be placed in a comparative context. The two most common contexts are comparisons of productivity for two different time periods for the same productive unit – e.g., for the same nation –

[13] Arguments for using other measures of total national output can be found, for example, in Kohli (1978, 1991, 2004, 2005, 2007). Diewert (2006d, 2007a) argues for the use of measures that are net of anticipated depreciation and obsolescence of capital assets. See also Diewert, Nakamura and Schreyer (2007).

[14] There are large literatures on measuring the various input volumes. On the labor input, see for example Ahmad et al. (2003), Baldwin, Maynard and Wong (2005), Baldwin et al. (2005), Bresnahan, Brynjolfsson and Hitt (2002), Nakamura (1995), Jorgenson and Fraumeni (1992), Jorgenson, Gollop and Fraumeni (1987), Tang and MacLeod (2005), and Triplett (1990, 1991). On the capital inputs see, for example, Diewert (1977, 1980a, 1983, 2001b, 2004a, 2004b, 2005a, 2005b, 2005c), Diewert and Lawrence (2000, 2005), Diewert, Mizobuchi and Nomura (2007), Diewert and Schreyer (2006), Diewert and Wykoff (2007), Hicks (1961), T.P. Hill (1999, 2000), Hulten (1986, 1990, 1992, 1996), Jorgenson (1963, 1980, 1989, 1995a, 1995b, 1996), and Schreyer (2001, 2005). See also Baldwin and Tanguay (2006), de Haan et al. (2005), Gu and Tang (2004), Harper (2004), Harper, Berndt and Wood (1989), Hayashi and Nomura (2005), R.J. Hill and T.P. Hill (2003), Inklaar, O'Mahony and Timmer (2005), Kuroda and Nomura (2004), Morrison (1988, 1999), Nomura (2004, 2005), Schreyer (2001, 2005), Timmer and van Ark (2005) and Triplett (1996).

or a contemporaneous comparison for two different productive units such as for Canada and the United States. Comparative productivity measures are sometimes referred to more specifically as productivity growth, or as relative productivity, measures, depending on the nature of the comparison.

Economists have tended to prefer the most comprehensive of possible productivity statistics: total factor productivity, designated commonly as TFP and defined as output divided by a measure of total input – i.e., a price weighted aggregate of the volumes of all of the inputs used in producing the designated output. In terms of the components of (2.3-2) above, we can represent TFP as follows:

$$\text{TFP} \equiv \frac{\text{GDP}}{\left[\begin{array}{c}\text{Total}\\ \text{input}\end{array}\right]} = (\text{A}). \tag{2.3-3}$$

Statistical agencies charged with producing productivity figures for nations are painfully aware that they do not manage to take account of *all* of the inputs used in producing the output of a nation. Thus official statistics agencies usually refer to the measures they compile, which are intended and used as approximations to TFP indexes, as multifactor productivity measures. These MFP measures can also be represented in terms of the components of the decomposition of GDP; i.e., we have

$$\text{MFP} \equiv \frac{\text{GDP}}{\left[\begin{array}{c}\text{Measured}\\ \text{input}\end{array}\right]} = \text{TFP} \times (\text{B}). \tag{2.3-4}$$

Labor productivity measures are far easier to compile than TFP and MFP type measures because the only input information needed is for the volume of labor used in producing the designated output. Labor productivity measures also have an especially transparent relationship to per capita GDP, which has given these productivity measures special public policy appeal.

One way of measuring the labor input is as an average wage weighted aggregate of the hours of work for different types of workers. The resulting weighted hours productivity measure can be specified as follows in terms of the components of per capita GDP given in (2.3-2):

$$\text{WHLP} \equiv \frac{\text{GDP}}{\left[\begin{array}{c}\text{Total}\\ \text{labor}\\ \text{input}\end{array}\right]} = \text{TFP} \times (\text{B}) \times (\text{C}) = \text{MFP} \times (\text{C}). \tag{2.3-5}$$

A simpler and more common way of measuring the labor input is as total hours of work (i.e., as the *un*weighted sum). The resulting hours labor productivity measure can be specified as:

$$\begin{aligned}\text{HLP} &\equiv \frac{\text{GDP}}{\left[\begin{array}{c}\text{Total}\\ \text{work}\\ \text{hours}\end{array}\right]} = \text{TFP} \times (\text{B}) \times (\text{C}) \times (\text{D}) = \text{MFP} \times (\text{C}) \times (\text{D})\\ &= \text{WHLP} \times (\text{D}).\end{aligned} \tag{2.3-6}$$

An even simpler way of measuring the labor input is as the number of workers. The resulting worker labor productivity measure is:

$$\mathrm{LP} \equiv \frac{\mathrm{GDP}}{\left[\begin{array}{c}\text{Number}\\ \text{of}\\ \text{workers}\end{array}\right]} = \mathrm{TFP} \times (\mathrm{B}) \times (\mathrm{C}) \times (\mathrm{D}) \times (\mathrm{E}) = \mathrm{MFP} \times (\mathrm{C}) \times (\mathrm{D}) \times (\mathrm{E})$$
$$= \mathrm{WHLP} \times (\mathrm{D}) \times (\mathrm{E}) = \mathrm{HLP} \times (\mathrm{E}). \tag{2.3-7}$$

As noted above, to be meaningfully interpreted, productivity measures must usually be placed in a comparative context. Productivity growth (or relative productivity) is evaluated by the ratio of the labor productivity, MFP or TFP measures for period (or production scenario) t versus s.

3. Four TFPG concepts in the N–M case

> "But even if we confine our attention to what is ordinarily called a commodity, such as 'wheat,' we find ourselves dealing with a composite commodity made up of winter wheat, spring wheat, of varying grades."
>
> [Paul A. Samuelson (1983, p. 130), Foundations of Economic Analysis]

Obviously, nations produce multiple outputs using multiple inputs. How can we measure the four concepts of TFPG introduced in Subsection 2.1 in general multiple input, multiple output production situations? This is the question explored in this section.

We begin by defining volume aggregates that are components of the Paasche, Laspeyres, and Fisher Ideal (referred to hereafter simply as Fisher) volume, price and TFPG indexes, and then give the formulas for these indexes. Törnqvist and implicit Törnqvist index numbers are also defined.

3.1. Price weighted volume aggregates

For a general N-input, M-output production process, the period t input and output price vectors are denoted by $w^t \equiv [w_1^t, \ldots, w_N^t]$ and $p^t \equiv [p_1^t, p_2^t, \ldots, p_M^t]$, while $x^t \equiv [x_1^t, \ldots, x_N^t]$ and $y^t \equiv [y_1^t, \ldots, y_M^t]$ denote the period t input and output volume vectors.

Nominal total cost C^t and revenue R^t can be viewed as price weighted volume aggregates of the micro data for the transactions, and are defined as follows for period s and t:

$$C^t \equiv \sum_{n=1}^{N} w_n^t x_n^t, \qquad R^t \equiv \sum_{m=1}^{M} p_m^t y_m^t, \tag{3.1-1}$$

$$C^s \equiv \sum_{n=1}^{N} w_n^s x_n^s \quad \text{and} \quad R^s \equiv \sum_{m=1}^{M} p_m^s y_m^s. \tag{3.1-2}$$

We also define four hypothetical volume aggregates.[15] The first two result from evaluating period t volumes using period s price weights:

$$\sum_{n=1}^{N} w_n^s x_n^t \quad \text{and} \quad \sum_{m=1}^{M} p_m^s y_m^t. \tag{3.1-3}$$

These aggregates are what the cost and revenue would have been if the period t inputs and outputs had been transacted at period s prices. In contrast, the third and fourth aggregates are sums of period s volumes evaluated using period t prices:

$$\sum_{n=1}^{N} w_n^t x_n^s \quad \text{and} \quad \sum_{m=1}^{M} p_m^t y_m^s. \tag{3.1-4}$$

These are what the cost and revenue would have been if the period s inputs had been purchased and the period s outputs had been sold at period t prices. No assumptions are involved in defining the hypothetical volume aggregates.[16]

3.2. The Paasche, Laspeyres and Fisher volume and price indexes

The Paasche (1874), Laspeyres (1871), and Fisher (1922, p. 234) output volume indexes can be defined, respectively, as follows using the volume aggregates given in (3.1-1)–(3.1-4):

$$Q_P \equiv \sum_{m=1}^{M} p_m^t y_m^t \Big/ \sum_{m=1}^{M} p_m^t y_m^s, \tag{3.2-1}$$

$$Q_L \equiv \sum_{m=1}^{M} p_m^s y_m^t \Big/ \sum_{m=1}^{M} p_m^s y_m^s, \quad \text{and} \tag{3.2-2}$$

$$Q_F \equiv (Q_P Q_L)^{(1/2)}. \tag{3.2-3}$$

[15] Formally, the first two of these can be obtained by deflating the period t nominal cost and revenue by a Paasche price index. The second two result from deflating the period t nominal cost and revenue by a Laspeyres price index. See Horngren and Foster (1987, Chapter 24, Part One) or Kaplan and Atkinson (1989, Chapter 9) for examples of this common accounting practice of controlling for price level change without mention of price indexes. See also Armitage and Atkinson (1990).

[16] Traditionally these aggregates were defined as weighted averages of volume and price relatives. A volume (price) relative for a good is the ratio of the volume (price) for that good in a specified period t to the volume (price) for that good in some comparison period s. One advantage of defining a volume (or price) index as a weighted average of relatives is that the relatives are unit free, making it clear that this is an acceptable way of incorporating even goods (prices) for which there is no generally accepted unit of measure. The equivalent definitions presented here are more convenient for establishing that each of these TFPG indexes is a measure of all four of the different concepts of TFPG introduced in Subsection 2.1.

Similarly, the Paasche, Laspeyres, and Fisher input volume indexes can be defined as:

$$Q_P^* \equiv \sum_{n=1}^{N} w_n^t x_n^t \Big/ \sum_{n=1}^{N} w_n^t x_n^s, \tag{3.2-4}$$

$$Q_L^* \equiv \sum_{n=1}^{N} w_n^s x_n^t \Big/ \sum_{n=1}^{N} w_n^s x_n^s, \quad \text{and} \tag{3.2-5}$$

$$Q_F^* \equiv \left(Q_P^* Q_L^*\right)^{(1/2)}. \tag{3.2-6}$$

Output and input volume indexes are all that are needed to define measures of the first and second concepts of TFPG. However, in order to specify measures of the third and fourth concepts for the multiple input, multiple output case, price indexes are needed too.

Price indexes can be constructed using any of the functional forms given for volume indexes simply by reversing the roles of the prices and volumes. Thus output and input price indexes for the Paasche, Laspeyres and Fisher formulas are given by:

$$P_P \equiv \sum_{m=1}^{M} p_m^t y_m^t \Big/ \sum_{m=1}^{M} p_m^s y_m^t, \tag{3.2-7}$$

$$P_P^* \equiv \sum_{n=1}^{N} w_n^t x_n^t \Big/ \sum_{n=1}^{N} w_n^s x_n^t, \tag{3.2-8}$$

$$P_L \equiv \sum_{m=1}^{M} p_m^t y_m^s \Big/ \sum_{m=1}^{M} p_m^s y_m^s, \tag{3.2-9}$$

$$P_L^* \equiv \sum_{n=1}^{N} w_n^t x_n^s \Big/ \sum_{n=1}^{N} w_n^s x_n^s, \tag{3.2-10}$$

$$P_F \equiv (P_P P_L)^{(1/2)}, \quad \text{and} \tag{3.2-11}$$

$$P_F^* \equiv \left(P_P^* P_L^*\right)^{(1/2)}. \tag{3.2-12}$$

A price index is defined to be the implicit counterpart of a volume index if the product rule (also called the product test or axiom) is satisfied.[17] This rule requires that the product of the volume and price indexes must equal the total cost ratio for input side indexes or the total revenue ratio for output side indexes.[18] Usually the implicit price index will not have the same functional form as the volume index it is associated with. For example, the Paasche price index is the implicit counterpart of a Laspeyres volume

[17] For more on the properties of direct versus implicit indexes, see Allen and Diewert (1981).

[18] The implicit price (volume) index corresponding to a given volume (price) index can always be derived by imposing the product test and solving for the price (volume) index that satisfies this rule. The product test is part of the axiomatic approach to the choice of an index number functional form that is reviewed in Section 4.

index, and the Laspeyres price index is the implicit counterpart of a Paasche volume index. The Fisher indexes are unusual in that the Fisher price index satisfies the product test rule when paired with a Fisher volume index.[19]

In defining and proving equalities for the measures of the four concepts of TFPG for a general multiple input, multiple output production situation, we use the following implications of the product rule. In particular, for the Paasche, Laspeyres and Fisher indexes, on the input side we have

$$Q_P^* \times P_L^* = Q_L^* \times P_P^* = Q_F^* \times P_F^* = C^t/C^s, \tag{3.2-13a}$$

and on the output side we have

$$Q_P \times P_L = Q_L \times P_P = Q_F \times P_F = R^t/R^s. \tag{3.2-13b}$$

3.3. TFPG measures for the N–M case

The traditional definition of a total factor productivity growth index in the index number literature is as a ratio of output and input volume indexes:

$$\text{TFPG} \equiv Q/Q^*. \tag{3.3-1}$$

Thus the Paasche, Laspeyres, and Fisher TFPG indexes can be defined using the Paasche, Laspeyres, and Fisher volume indexes. Given a choice of *any one* of these three functional forms, we prove here that the corresponding multiple input, multiple output case measures are all equal for the four concepts of TFPG introduced in Subsection 2.1.

We proceed as follows to establish these equalities for the measures of the TFPG(1), TFPG(2) and TFPG(3) concepts. We first use the product rule results to define Paasche, Laspeyres and Fisher TFPG(3) measures. We substitute in the definitions of the components of the TFPG(3) measures and rearrange terms to establish the equalities with the TFPG(2) and TFPG(1) measures. Then we take up the TFPG(4) case.

For a Paasche TFPG index we have:

$$\begin{aligned}\text{TFPG}_P &= \frac{Q_P}{Q_P^*} = \frac{(R^t/R^s)/P_L}{(C^t/C^s)/P_L^*} \equiv \text{TFPG(3)}_P \quad \text{using (3.3-1) and (3.2-13)} \\ &= \frac{\sum_{m=1}^M p_m^t y_m^t / \sum_{m=1}^M p_m^t y_m^s}{\sum_{n=1}^N w_n^t x_n^t / \sum_{n=1}^N w_n^t x_n^s} \equiv \text{TFPG(2)}_P \\ &\qquad \text{using (3.1-1), (3.1-2) and also (3.2-9) and (3.2-10)}\end{aligned}$$

[19] When the product of a price and a volume index that both have the same formula equals the value ratio (i.e., the revenue ratio in the case of output indexes, or the cost ratio in the case of input indexes), then the formula satisfies the factor reversal test. The Fisher formula is unusual, but not unique, in satisfying this test. See Diewert (1987) on the factor reversal test.

$$= \frac{\sum_{m=1}^{M} p_m^t y_m^t / \sum_{n=1}^{N} w_n^t x_n^t}{\sum_{m=1}^{M} p_m^t y_m^s / \sum_{n=1}^{N} w_n^t x_n^s} \equiv \text{TFPG}(1)_P. \tag{3.3-2}$$

For a Laspeyres TFPG index we have:

$$\text{TFPG}_L = \frac{Q_L}{Q_L^*} = \frac{(R^t/R^s)/P_P}{(C^t/C^s)/P_P^*} \equiv \text{TFPG}(3)_L \quad \text{using (3.3-1) and (3.2-13)}$$

$$= \frac{\sum_{m=1}^{M} p_m^s y_m^t / \sum_{m=1}^{M} p_m^s y_m^s}{\sum_{n=1}^{N} w_n^s x_n^t / \sum_{n=1}^{N} w_n^s x_n^s} \equiv \text{TFPG}(2)_L$$

using (3.1-1), (3.1-2) and also (3.2-7) and (3.2-8)

$$= \frac{\sum_{m=1}^{M} p_m^s y_m^t / \sum_{n=1}^{N} w_n^s x_n^t}{\sum_{m=1}^{M} p_m^s y_m^s / \sum_{n=1}^{N} w_n^s x_n^s} \equiv \text{TFPG}(1)_L. \tag{3.3-3}$$

And for a Fisher TFPG index we have:

$$\text{TFPG}_F = \frac{Q_F}{Q_F^*} = \frac{(R^t/R^s)/P_F}{(C^t/C^s)/P_F^*} \equiv \text{TFPG}(3)_F \quad \text{using (3.3-1) and (3.2-13)}$$

$$= \frac{[(\frac{R^t}{R^s})P_L]^{1/2}[(\frac{R^t}{R^s})P_P]^{1/2}}{[(\frac{C^t}{C^s})P_L^*]^{1/2}[(\frac{C^t}{C^s})P_P^*]^{1/2}} = \frac{\left[\frac{\sum_{m=1}^{M} p_m^t y_m^t}{\sum_{m=1}^{M} p_m^t y_m^s}\right]^{1/2}\left[\frac{\sum_{m=1}^{M} p_m^s y_m^t}{\sum_{m=1}^{M} p_m^s y_m^s}\right]^{1/2}}{\left[\frac{\sum_{n=1}^{N} w_n^t x_n^t}{\sum_{n=1}^{N} w_n^t x_n^s}\right]^{1/2}\left[\frac{\sum_{n=1}^{N} w_n^s x_n^t}{\sum_{n=1}^{N} w_n^s x_n^s}\right]^{1/2}}$$

$$\equiv \text{TFPG}(2)_F$$

using (3.2-3), (3.2-13), (3.1-1), (3.1-2), and (3.2-7)–(3.2-10)

$$= \frac{\left[\frac{\sum_{m=1}^{M} p_m^t y_m^t}{\sum_{n=1}^{N} w_n^t x_n^t}\right]^{1/2}\left[\frac{\sum_{m=1}^{M} p_m^s y_m^t}{\sum_{n=1}^{N} w_n^s x_n^t}\right]^{1/2}}{\left[\frac{\sum_{m=1}^{M} p_m^t y_m^s}{\sum_{n=1}^{N} w_n^t x_n^s}\right]^{1/2}\left[\frac{\sum_{m=1}^{M} p_m^t y_m^s}{\sum_{n=1}^{N} w_n^t x_n^s}\right]^{1/2}} \equiv \text{TFPG}(1)_F. \tag{3.3-4}$$

The TFPG(4) concept is the rate of growth in the margin after controlling for price change. In the N–M case, just as in the 1–1 one, the margin m^t is given by

$$1 + m^t \equiv R^t/C^t. \tag{3.3-5}$$

Depending on whether Laspeyres, Paasche or Fisher price indexes are used to deflate the cost and revenue components of the margin, the respective expressions for TFPG(3) given in (3.3-2), (3.3-3) and (3.3-4) can be rewritten as:

$$\text{TFPG}(4)_P \equiv \left[\left(1+m^t\right)/\left(1+m^s\right)\right]\left[P_L^*/P_L\right], \tag{3.3-6}$$

$$\text{TFPG}(4)_L \equiv \left[\left(1+m^t\right)/\left(1+m^s\right)\right]\left[P_P^*/P_P\right], \quad \text{and} \tag{3.3-7}$$

$$\text{TFPG}(4)_F \equiv \left[\left(1+m^t\right)/\left(1+m^s\right)\right]\left[P_F^*/P_F\right]. \tag{3.3-8}$$

Notice that if the margins are zero, regardless of the reasons, then each of these expressions for TFPG(4) reduces to the ratio of the input price index to the output price index.[20]

3.4. Other index number formulas

Many other index number formulas have been proposed besides the Paasche, Laspeyres and Fisher.[21] Here we will use Q_G and P_G and Q_G^* and P_G^* to denote any two pairs of direct and implicit output and input volume and price indexes. These are any output side and input side pairs of volume and price indexes that satisfy the product rule so that $Q_G P_G = (R^t/R^s)$ and $Q_G^* P_G^* = (C^t/C^s)$. From these product rule results and (3.3-5), it is easily seen that the following measures of concepts (3.3-2), (3.3-3) and (3.3-4) of TFPG are all equal:

$$\begin{aligned}\frac{(R^t/R^s)/P_G}{(C^t/C^s)/P_G^*} &\equiv \text{TFPG}(3)_G \\ &= Q_G/Q_G^* \equiv \text{TFPG}(2)_G \\ &= \left[\left(1+m^t\right)/\left(1+m^s\right)\right]\left[P^*/P\right] \equiv \text{TPFG}(4). \end{aligned} \tag{3.4-1}$$

This is a general result that nests the results given in Subsection 3.3.

But what about $\text{TFPG}(1)_G$? A measure of the growth in the rate of transformation of total input into total output ideally should be defined using measures of input and output that are comparable for period s and t in the sense that the micro level volumes for both periods are aggregated using the same price weights. This is a desirable property if levels comparisons are to be made for pairs of nations. The volume aggregates that are the components of the Paasche, Laspeyres and Fisher TFPG(1) measures defined in the first line of (3.3-2), (3.3-3) and (3.3-4) satisfy what we refer to as this *comparability over time ideal*.[22] There are many other index number formulas for which it is not possible to define this sort of a measure for the TFPG(1) concept that also equals the corresponding measures for the other three concepts of TFPG. For those that are nevertheless superlative, an approximate equality of $\text{TFPG}(1)_G$ with the expressions for the other three concepts of TFPG is established as follows.

[20] One set of conditions under which the margins will be zero is perfect competition and a constant returns to scale technology.

[21] See Diewert (1993b, 1993c) and Fisher (1911, 1922).

[22] The period t cost and revenue and the hypothetical aggregates of period s output and input volumes defined in expressions (3.1-1) and (3.1-4) are comparable in this sense because the volumes for period s and t are evaluated using the same period t price vectors. Similarly, the period s cost and revenue and the hypothetical aggregates of period t output and input volumes defined in expressions (3.1-2) and (3.1-3) are comparable in this sense because the volumes of the output and input goods are evaluated using the same period s price vectors. These aggregates are what are used to define the Paasche, Laspeyres and Fisher measures given in (3.3-2), (3.3-3) and (3.3-4).

For any pair of volume and price indexes satisfying the product test, from (3.4-1) and the product rule implications we see that the following expressions equal, respectively, those given in (3.4-1) for TFPG(2)$_G$, TFPG(3)$_G$ and TFPG(4)$_G$:

$$\frac{Q_G}{Q_G^*} = \frac{(R^t/R^s)/P}{(C^t/C^s)/P^*} = \frac{\sum_{m=1}^{M}(p_m^t/P_G)y_m^t/\sum_{m=1}^{M}p_m^s y_m^s}{\sum_{n=1}^{N}(w_n^t/P_G^*)x_n^t/\sum_{n=1}^{N}w_n^s x_n^s}. \tag{3.4-2}$$

In the last of these expressions, the price vectors (p^t/P_G) and (w^t/P_G^*) appearing in the period t output and input volume aggregates are the period t prices expressed in period s dollars. If we choose this expression as the measure of TFPG(1)$_G$, then with the choice of a Paasche, Laspeyres or Fisher formula, this measure will be ideal in the sense of using the same price weights to compare the period t and s volumes. When some other formula is used, there is an approximate solution to this problem for indexes that satisfy the product rule and are also "superlative". This approximate solution makes use of the Fisher functional form with the TFPG(1) measure, defined as in the last line of (3.3-4).

Diewert coined the term superlative for an index number functional form that is "exact" in that it can be derived algebraically from a producer or consumer behavioral equation that satisfies the Diewert flexibility criterion. According to this criterion, a functional form is flexible if it can provide a second order approximation to an arbitrary twice continuously differentiable linearly homogeneous function. Diewert (1976, 1978b) established that under usual conditions, all of the commonly used superlative index number formulas (including the Fisher, and also the Törnqvist and implicit Törnqvist functional forms introduced below) approximate each other to the second order when evaluated at an equal price and volume point. This is a numerical analysis approximation result that does not rely on any further assumptions.[23]

Because the Fisher volume and price indexes satisfy the product rule, we have

$$Q_G P_G = \left(R^t/R^s\right) = Q_F P_F \quad \text{and} \quad Q_G^* P_G^* = \left(C^t/C^s\right) = Q_F^* P_F^*,$$

and dividing through by P_G and P_G^*, respectively, yields

$$\frac{Q_G}{Q_G^*} = \left[\frac{Q_F}{Q_F^*}\right]\left[\frac{P_F/P_G}{P_F^*/P_G^*}\right]. \tag{3.4-3}$$

From (3.4-3), (3.4-1) and (3.3-4) we see that if we define the measure for the first concept of TFPG as

$$\text{TPFG(1)}_G \equiv \text{TPFG(1)}_F\left[\frac{P_F/P_G}{P_F^*/P_G^*}\right], \tag{3.4-4}$$

this measure will equal TFPG(2)$_G$, TFPG(3)$_G$ and TFPG(4)$_G$ as defined in (3.4-1). In this TFPG(1)$_G$ measure, the period t price vectors, p^t and w^t, of the TFPG(1)$_F$ component are replaced by $(p^t/(P_F/P_G))$ and $(w^t/(P_F^*/P_G^*))$. As a consequence, unless

[23] R.J. Hill (2006) shows, however, that being superlative does not, by itself, ensure an index is desirable.

the given price indexes are Laspeyres or Paasche or Fisher ones, the period t and s volumes compared by the measure will not be aggregated using the same price weights when there have been changes in relative prices. Nevertheless, for superlative index numbers, it follows that when the chosen volume and price indexes are any of the commonly used ones such as the Törnqvist or implicit Törnqvist, then we can use the result that, under usual conditions, all of the superlative indexes in common use approximate each other to the second order at an equal price and volume point. That is, we have $\text{TFPG}(1)_G \cong \text{TFPG}(1)_F$.

3.5. The Törnqvist (or Translog) indexes[24]

Törnqvist (1936) indexes are weighted geometric averages of growth rates for the volume or price relatives for the different products. These indexes have been widely used by statistical agencies and in the economics literature. The formula for the natural logarithm of a Törnqvist index is usually shown as the definition for this index. For the output volume index, this is

$$\ln Q_T = (1/2)\sum_{m=1}^{M}\left[\left(p_m^s y_m^s \Big/ \sum_{i=1}^{M} p_i^s y_i^s\right) + \left(p_m^t y_m^t \Big/ \sum_{j=1}^{M} p_j^t y_j^t\right)\right] \ln\left(y_m^t / y_m^s\right). \tag{3.5-1}$$

The Törnqvist input volume index Q_T^* is defined analogously, with input volumes and prices substituted for the output volumes and prices in (3.5-1).

Reversing the role of the prices and volumes in the formula for the Törnqvist output volume index yields the Törnqvist output price index, P_T, defined by

$$\ln P_T = (1/2)\sum_{m=1}^{M}\left[\left(p_m^s y_m^s \Big/ \sum_{i=1}^{M} p_i^s y_i^s\right) + \left(p_m^t y_m^t \Big/ \sum_{j=1}^{M} p_j^t y_j^t\right)\right] \ln\left(p_m^t / p_m^s\right). \tag{3.5-2}$$

The input price index P_T^* is defined in a similar manner.

The implicit Törnqvist output volume index, denoted by $Q_{\widetilde{T}}$, is defined implicitly by[25] $(R^t/R^s)/P_T \equiv Q_{\widetilde{T}}$, and the implicit Törnqvist input volume index, $Q_{\widetilde{T}}^*$, is defined analogously using the cost ratio and P_T^*. The implicit Törnqvist output price index, $P_{\widetilde{T}}$, is given by $(R^t/R^s)/Q_T \equiv P_{\widetilde{T}}$, and the implicit Törnqvist input price index, $P_{\widetilde{T}}^*$, is defined analogously.

Using the Törnqvist volume and the implicit Törnqvist price indexes, or the implicit Törnqvist volume and the Törnqvist price indexes, measurement formulas for the second, third and fourth concepts of TFPG can be specified as in (3.4-1) above. Moreover,

[24] Törnqvist indexes are also known as translog indexes following Jorgenson and Nishimizu (1978) who introduced this terminology because Diewert (1976, p. 120) related Q_T^* to a translog production function. The exact index number approach used for relating specific volume indexes to specific production functions is the topic of Section 5.

[25] See Diewert (1992a, p. 181).

these are superlative indexes for which Section 3.4 approximation result applies; that is, we have $\text{TFPG}(1)_T \cong \text{TFPG}(1)_F$ and $\text{TFPG}(1)_{\widetilde{T}} \cong \text{TFPG}(1)_F$.

4. The axiomatic (or test) approach to index formula choice

Multiple TFPG index number formulas can all be viewed as measures of total factor productivity growth. This was demonstrated in Section 3 for the commonly used Laspeyres, Paasche, Fisher and Törnqvist indexes, and this result could be established for other proposed index number formulas as well. Since different formulas will yield different estimates for TFPG, which one should be used, and why? Historically, index number theorists have relied on what is called the axiomatic or test approach to address this functional form choice problem. An overview of this approach is provided here.

As before, Q denotes an output volume index and P denotes an output price index. The corresponding input volume and price indexes are denoted by the same symbols with a star superscript added. The axiomatic approach to the determination of the functional forms for Q and P on the output side, or for Q^* and P^* on the input side, works as follows. The starting point is a list of mathematical properties that *a priori* reasoning suggests a price index should satisfy. These are the index number theory 'tests' or 'axioms'. Mathematical reasoning is applied to determine whether the *a priori* tests are mutually consistent and whether they uniquely determine, or usefully narrow, the choice of the functional form for the price index.[26] Once the form of the price index has been decided on, imposition of the product test rule determines the functional form of the volume index as well.

The *product test* was already introduced in Subsection 3.2.[27] On the output side, this rule states that the product of the output price and output volume indexes, P and Q, should equal the nominal revenue ratio for periods t and s:

$$PQ = R^t/R^s. \tag{4-1}$$

If the functional form for the output price index P is given, then imposing the product rule means that the functional form for the volume index must be given by the expression[28]

$$Q = \left(R^t/R^s\right)/P. \tag{4-2}$$

[26] Contributors to this approach include Walsh (1901, 1921), Fisher (1911, 1922), Eichhorn (1976), Eichhorn and Voeller (1976), Funke and Voeller (1978, 1979), Diewert (1976, 1987, 1988, 1992b, 1999), Balk (1995) and Armstrong (2003).

[27] The product test was proposed by Irving Fisher (1911, p. 388) and named by Frisch (1930, p. 399).

[28] Volume or price indexes derived by imposing the product rule and specifying the form of the price or volume index are sometimes referred to as implicit indexes. The $\sim$ symbol is sometimes added on top of the symbol for the index number when it is desired to call attention to the implicit nature of the index. Any test that satisfies the factor reversal test would also satisfy the product test.

Thus, unlike the other tests introduced below that are applied to the alternative price indexes of interest and that may be passed or failed by each of the index number formulas tested, the product test is often imposed at the outset as part of the formula choice process.[29]

We conclude this overview of the axiomatic approach by listing four tests that can be applied for choosing among alternative functional forms for the price index. Only the output side price indexes are considered here, but the tests are applied in the same manner on the input side.

The *identity* or *constant prices test* is[30]

$$P\left(p, p, y^s, y^t\right) = 1. \tag{4-3}$$

What this means is that if all prices stay the same over the current and comparison time periods so that $p^s = p^t = p = (p_1, \ldots, p_M)$, then the price index should be one regardless of the volume values for period s and t.

The *constant basket test*, also called the *constant volumes test*, is[31]

$$P\left(p^s, p^t, y, y\right) = \sum_{i=1}^{N} p_i^t y_i \Big/ \sum_{j=1}^{N} p_j^s y_j. \tag{4-4}$$

This test states that if the volumes produced for all output goods stay the same for period s and t so that $y^s = y^t = y \equiv (y_1, \ldots, y_M)$, then the level of prices in period t compared to s should equal the value of the constant basket of volumes evaluated at the period t prices divided by the value of this same basket evaluated at the period s prices.

The *proportionality in period t prices test* is[32]

$$P\left(p^s, \lambda p^t, y^s, y^t\right) = \lambda P\left(p^s, p^t, y^s, y^t\right) \quad \text{for } \lambda > 0. \tag{4-5}$$

According to this test, if each of the elements of p^t is multiplied by the positive constant λ, then the level of prices in period t relative to s should differ by the same multiplicative factor λ.

Our final example of a price index test is the *time reversal test*[33]:

$$P\left(p^t, p^s, y^t, y^s\right) = 1/P\left(p^s, p^t, y^s, y^t\right). \tag{4-6}$$

[29] Note that the product test is not the same as the factor reversal test, although any formula that satisfies the factor reversal test will satisfy the product test. As pointed out to us by Andy Baldwin in private correspondance, in imposing the product test on a price index, one normally has already chosen the volume index and the price index is chosen by default to satisfy the product index. Thus the Paasche formula is chosen for the price index because one would like to have a Laspeyres volume index.

[30] This test was proposed by Laspeyres (1871, p. 308), Walsh (1901, p. 308) and Eichhorn and Voeller (1976, p. 24).

[31] This test was proposed by many researchers including Walsh (1901, p. 540).

[32] This test was proposed by Walsh (1901, p. 385) and Eichhorn and Voeller (1976, p. 24).

[33] This test was first informally proposed by Pierson (1896, p. 128) and was formalized by Walsh (1901, p. 368, 1921, p. 541) and Fisher (1922, p. 64).

If this test is satisfied, then when the prices and volumes for period s and t are interchanged, the resulting price index will be the reciprocal of the original price index.

The Paasche and Laspeyres indexes, P_P and P_L, fail the time reversal test (4-6). The Törnqvist index, P_T, fails the constant basket test (4-4), and the implicit Törnqvist index, $\widetilde{P}_T$, fails the constant prices test (4-3). On the other hand, the Fisher price index P_F satisfies all four of these tests. When a more extensive list of tests is compiled, the Fisher price index continues to satisfy more tests than other leading candidates.[34] These results favor the Fisher TFPG index. However, the Paasche, Laspeyres, Törnqvist, and implicit Törnqvist indexes all rate reasonably well according to the axiomatic approach.

5. The exact approach and superlative index numbers

> "Tinbergen (1942, pp. 190–195) interprets the geometric volume index of total factor productivity as a Cobb–Douglas production function. As further examples of index-number formulas that have been interpreted as production functions, a fixed-weight Laspeyres volume index of total factor productivity may be interpreted as a 'linear' production function, that is, as a production function with infinite elasticity of substitution, as Solow (1957, p. 317) and Clemhout (1963, pp. 358–360) have pointed out. In a sense, output-capital or output-labor ratios correspond to Leontief-type production functions, that is, to production functions with zero elasticity of substitution, as Domar (1961, pp. 712–713) points out."
>
> [Dale W. Jorgenson (1995a, p. 48), Productivity Vol. 1]

An alternative approach to the determination of the functional form for a measure of total factor productivity growth is to derive the TFPG index from a producer behavioral model. Diewert's (1976) exact index number approach is a paradigm for doing this. This approach places the index number formula choice problem on familiar territory for economists, allowing the choice to be based on axioms of economic behavior or empirical evidence about producer behavior rather than, or in addition to, the traditional tests of the axiomatic approach.

The exact index number approach is perhaps most easily explained by outlining the main steps in an actual application. In this section we sketch the steps involved in deriving a TFPG index that is exact for a translog cost function for which certain stated restrictions hold.

The technology of a firm can be summarized by its period t production function f^t. If we focus on the production of output 1, then the period t production function can be represented as

$$y_1 = f^t(y_2, y_3, \ldots, y_M, x_1, x_2, \ldots, x_N). \tag{5-1}$$

[34] See Diewert (1976, p. 131, 1992b) and also Funke and Voeller (1978, p. 180).

This function gives the amount of output 1 the firm can produce using the technology available in any given period t if it also produces y_m units of each of the outputs $m = 2, \ldots, M$ using x_n units for each of the inputs $n = 1, \ldots, N$.

The production function f^t can be used to define the period t cost function:

$$c^t(y_1, y_2, \ldots, y_M, w_1, w_2, \ldots, w_N). \tag{5-2}$$

This function is postulated to give the minimum cost of producing the output volumes $y_1, \ldots, y_M$ using the period t technology and with the given input prices w_n^t, $n = 1, 2, \ldots, N$. Under the assumption of cost minimizing behavior, the observed period t cost of production, denoted by C^t, is the minimum possible cost, and we have

$$C^t \equiv \sum_{n=1}^{N} w_n^t x_n^t = c^t\left(y_1^t, \ldots, y_M^t, w_1^t, \ldots, w_N^t\right). \tag{5-3}$$

We need some way of relating the cost functions for different time periods (or scenarios) to each other. One way is to assume the cost function for each period is a period specific multiple of an atemporal cost function. As a simplest (and much used case), we might assume that

$$c^t(y_1, \ldots, y_M, w_1, \ldots, w_N) = \left(1/a^t\right)c(y_1, \ldots, y_M, w_1, \ldots, w_N), \quad t = 0, 1, \ldots, T, \tag{5-4}$$

where $a^t > 0$ denotes a period t relative efficiency parameter and c denotes an atemporal cost function which does not depend on time. We have assumed in (5-4) that technological change is Hicks neutral. The normalization $a^0 \equiv 1$ is usually imposed. Given (5-4), a natural measure of productivity change (or relative productivity) for a productive unit for period t versus s is the ratio

$$a^t/a^s. \tag{5-5}$$

If this ratio is greater than 1, efficiency is said to have improved.

Taking the natural logarithm of both sides of (5-4), we have

$$\ln c^t\left(y_1^t, \ldots, y_M^t, w_1^t, \ldots, w_N^t\right) = -\ln a^t + \ln c\left(y_1^t, \ldots, y_M^t, w_1^t, \ldots, w_N^t\right). \tag{5-6}$$

Suppose that *a priori* information is available indicating that a translog functional form is appropriate for $\ln c$. In this case, the atemporal cost function c on the right-hand side of (5-6) can be represented by

$$\begin{aligned}\ln c\left(y_1^t, \ldots, y_M^t, w_1^t, \ldots, w_N^t\right) &= b_0 + \sum_{m=1}^{M} b_m \ln y_m^t + \sum_{n=1}^{N} c_n \ln w_n^t + (1/2)\sum_{i=1}^{M}\sum_{j=1}^{M} d_{ij} \ln y_i^t \ln y_j^t \\ &\quad + (1/2)\sum_{n=1}^{N}\sum_{j=1}^{N} f_{nj} \ln w_n^t \ln w_j^t + \sum_{m=1}^{M}\sum_{n=1}^{N} g_{mn} \ln y_m^t \ln w_n^t.\end{aligned} \tag{5-7}$$

An advantage of the choice of the translog functional form for the atemporal cost function part of (5-6) is that it does not impose *a priori* restrictions on the admissible patterns of substitution between inputs and outputs, but this flexibility results from a large number of free parameters.[35] There are $M+1$ of the b_m parameters, N of the c_n parameters, MN of the g_{mn} parameters, $M(M+1)/2$ independent d_{ij} parameters and $N(N+1)/2$ independent f_{nj} parameters even when it is deemed reasonable to impose the symmetry conditions that $d_{ij} = d_{ji}$ for $1 \leqslant i < j \leqslant M$ and $f_{nj} = f_{jn}$ for $1 \leqslant n < j \leqslant N$. If homogeneity of degree one in the input prices is also assumed, then the following additional restrictions hold for the parameters of (5-7):

$$\sum_{n=1}^{N} c_n = 1, \quad \sum_{j=1}^{N} f_{nj} = 0 \quad \text{for } n = 1, \ldots, N, \quad \text{and}$$

$$\sum_{n=1}^{N} g_{mn} = 0 \quad \text{for } m = 1, \ldots, M. \tag{5-8}$$

With all of the above restrictions, the number of independent parameters in (5-6) and in (5-7) is still $T + M(M+1)/2 + N(N+1)/2 + MN$. The number of parameters can easily end up being larger than the number of available observations.[36] Thus, without imposing more restrictions, it may not be possible to reliably estimate the parameters of (5-6) or to derive a productivity index from this sort of an estimated relationship.

One way of proceeding is to assume the producer is minimizing costs so that the following demand relationships hold[37]:

$$x_n^t = \partial c^t\left(y_1^t, \ldots, y_M^t, w_1^t, \ldots, w_N^t\right)/\partial w_n$$
$$\text{for } n = 1, \ldots, N \text{ and } t = 0, 1, \ldots, T. \tag{5-9}$$

Since $\ln c^t$ can also be regarded as a quadratic function in the variables

$$\ln y_1, \ln y_2, \ldots, \ln y_M, \ln w_1, \ln w_2, \ldots, \ln w_N,$$

[35] The translog functional form for a single output technology was introduced by Christensen, Jorgenson and Lau (1971, 1973). See also Christensen and Jorgenson (1973). The multiple output case was defined by Burgess (1974) and Diewert (1974a, p. 139).

[36] On the econometric estimation of cost and related aggregator functions using more flexible functional forms that permit theoretically plausible types of substitution, see for example Berndt (1991), Berndt and Khaled (1979) and also Diewert (1969, 1971, 1973, 1974b, 1978a, 1981a, 1982) and Diewert and Wales (1992, 1995).

[37] This follows by applying a theoretical result due initially to Hotelling (1925) and Shephard (1953, p. 11).

Diewert's (1976, p. 119) logarithmic quadratic identity can be applied. Accordingly, we have[38]:

$$\begin{aligned}\ln c^t - \ln c^s &= (1/2)\sum_{m=1}^{M}\left[y_m^t \frac{\partial \ln c^t}{\partial y_m}\left(y^t, w^t\right) + y_m^s \frac{\partial \ln c^s}{\partial y_m}\left(y^s, w^s\right)\right]\ln\left(y_m^t/y_m^s\right)\\ &\quad + (1/2)\sum_{n=1}^{N}\left[w_n^t \frac{\partial \ln c^t}{\partial w_n}\left(y^t, w^t\right) + w_n^s \frac{\partial \ln c^s}{\partial w_n}\left(y^s, w^s\right)\right]\ln\left(w_n^t/w_n^s\right)\\ &\quad + (1/2)\left[\frac{\partial \ln c^t}{\partial a}\left(y^t, w^t\right) + \frac{\partial \ln c^s}{\partial a}\left(y^s, w^s\right)\right]\ln\left(a^t/a^s\right) \qquad (5\text{-}10)\\ &= (1/2)\sum_{m=1}^{M}\left[y_m^t \frac{\partial \ln c^t}{\partial y_m}\left(y^t, w^t\right) + y_m^s \frac{\partial \ln c^s}{\partial y_m}\left(y^s, w^s\right)\right]\ln\left(y_m^t/y_m^s\right)\\ &\quad + (1/2)\sum_{n=1}^{N}\left[\left(w_n^t x_n^t/C^t\right) + \left(w_n^s x_n^s/C^s\right)\right]\ln\left(w_n^t/w_n^s\right)\\ &\quad + (1/2)\left[-1 + (-1)\right]\ln\left(a^t/a^s\right). \qquad (5\text{-}11)\end{aligned}$$

If it is acceptable to impose the additional assumption of competitive profit maximizing behavior, we can simplify (5-11) even further. More specifically, suppose we can assume that the output volumes $y_1^t, \ldots, y_M^t$ solve the following profit maximization problem for $t = 0, 1, \ldots, T$:

$$\max_{y_1,\ldots,y_M}\left[\sum_{m=1}^{M} p_m^t y_m - c^t\left(y_1, \ldots, y_M, w_1^t, \ldots, w_N^t\right)\right]. \qquad (5\text{-}12)$$

This leads to the usual price equals marginal cost relationships that result when competitive price taking behavior is assumed; i.e., we now have

$$p_m^t = \partial c^t(y_1^t, \ldots, y_M^t, w_1^t, \ldots, w_N^t)/\partial y_m, \quad m = 1, \ldots, M. \qquad (5\text{-}13)$$

This key step permits the use of observed prices as weights for aggregating the observed volume data for the different outputs and inputs. Making use of the definition of total costs in (5-3), expression (5-11) can now be rewritten as:

$$\begin{aligned}\ln\left(C^t/C^s\right) &= (1/2)\sum_{m=1}^{M}\left[\left(p_m^t y_m^t/C^t\right) + \left(p_m^s y_m^s/C^s\right)\right]\ln\left(y_m^t/y_m^s\right)\\ &\quad + (1/2)\sum_{n=1}^{N}\left[\left(w_n^t x_n^t/C^t\right) + \left(w_n^s x_n^s/C^s\right)\right]\ln\left(w_n^t/w_n^s\right) - \ln\left(a^t/a^s\right).\end{aligned} \qquad (5\text{-}14)$$

[38] Expression (5-11) follows from (5-10) by applying the Hotelling–Shephard relations (5-9) for period t and s.

Total costs in period s and t presumably can be observed, as can output and input prices and volumes. Thus the only unknown in Equation (5-14) is the productivity change measure going from period s to t. Solving (5-14) for this measure yields

$$a^t/a^s = \left\{ \prod_{m=1}^{M} \left(y_m^t/y_m^s\right)^{(1/2)[(p_m^t y_m^t/C^t)+(p_m^s y_m^s/C^s)]} \right\} / \widetilde{Q}_T^*, \tag{5-15}$$

where $\widetilde{Q}_T^*$ is the implicit Törnqvist input volume index that is defined analogously to the implicit Törnqvist output volume index introduced in Subsection 3.5.

Formula (5-15) can be simplified still further if it is appropriate to assume that the underlying technology exhibits constant returns to scale. If costs grow proportionally with output, then it can be shown [e.g., see Diewert (1974a, pp. 134–137)] that the cost function must be linearly homogeneous in the output volumes. In that case, with competitive profit maximizing behavior, revenues must equal costs in each period. In other words, under the additional hypothesis of constant returns to scale, for each time period $t = 0, 1, \ldots, T$ we have the equality:

$$c^t\left(y^t, w^t\right) = C^t = R^t. \tag{5-16}$$

Using (5-16), we can replace C^t and C^s in (5-15) by R^t and R^s, and (5-15) becomes

$$a^t/a^s = Q_T/\widetilde{Q}_T^*, \tag{5-17}$$

where Q_T is the Törnqvist output volume index and $\widetilde{Q}_T^*$ is the implicit Törnqvist input volume index. This means that if we can justify the choice of a translog cost function and if the assumptions underlying the above derivations are true, then we have a basis for choosing $(Q_T/\widetilde{Q}_T^*)$ as the appropriate functional form of the TFPG index.

The hypothesis of constant returns to scale that must be invoked in moving from expression (5-15) to (5-17) is very restrictive. However, if the underlying technology is subject to diminishing returns to scale, we can convert the technology into an artificial one still subject to constant returns to scale by introducing an extra fixed input, x_{N+1} say, and setting this extra fixed input equal to one (that is, $x_{N=1}^t = 1$ for each period t). The corresponding period t price for this input, w_{N+1}^t, is set equal to the firm's period t profits, $R^t - C^t$. With this extra factor, the firm's period t cost is redefined to be the adjusted cost given by

$$C_A^t = C^t + w_{N+1}^t x_{N+1}^t = \sum_{n=1}^{N+1} w_n^t x_n^t = R^t. \tag{5-18}$$

The derivation can now be repeated using the adjusted cost C_A^t rather than the actual cost C^t. This results in the same productivity change formula except that $\widetilde{Q}_T^*$ is now the implicit translog volume index for $N + 1$ instead of N inputs. Thus, in the diminishing returns to scale case, we could use formula (5-15) as our measure of productivity change between period s and t, or we could use formula (5-17) with the understanding that the

extra fixed input would then be added into the list of inputs and incorporated into the adjusted costs.

Formulas (5-15) and (5-17) illustrate the exact index number approach to the derivation of productivity change measures. The method may be summarized as follows: (1) *a priori* or empirical evidence is used as a basis for choosing a specific functional form for the firm's cost function,[39] (2) competitive profit maximizing behavior is assumed (or else cost minimizing plus competitive revenue maximizing behavior), and (3) various identities are manipulated and a productivity change measure emerges that depends only on observable prices and volumes.

In this section, the use of the exact index number method has been demonstrated for a situation where the functional form for the cost function was assumed to be adequately approximated by a translog with parameters satisfying symmetry, homogeneity, cost minimization, profit maximization, and possibly also constant returns to scale conditions. The resulting productivity change term a^t/a^s given by the formula on the right-hand side of (5-15) or (5-17) can be directly evaluated even with thousands of outputs and inputs.

It is important to bear in mind, however, that all of the index number TFPG measures defined in Section 3 can be evaluated numerically for each time period given suitable volume and price data regardless of whether assumptions such as those made above are true. The assumptions are used only to show that particular TFPG index number formulas can be derived from certain optimizing models of producer behavior. Such a model might then be used in interpreting the TFPG value. For instance, the model might be used as a basis for breaking up the TFPG value into returns to scale and technical progress components. Decompositions of this sort are taken up in Sections 6.1, 10.2 and 11.

6. Production function based measures of TFPG

When a TFPG index can be related to a producer behavioral relationship that is derived from an optimizing model of producer behavior, this knowledge provides a potential theoretical basis for defining various decompositions of TFPG and interpreting component parts. This is the approach adopted here.

We begin in Subsection 6.1 by considering some production function based alternatives for factoring TFPG into technical progress (TP) and returns to scale (RS) components in the simplified one input, one output case. Even in the general multiple input, multiple output case, a TP and RS decomposition of TFPG has no direct implications for the choice of a measurement formula for TFPG since the new parameters introduced

[39] In place of step (1) where a specific functional form is assumed for the firm's cost function, some researchers have specified functional forms for the firm's production function [e.g., Diewert (1976, p. 127)] or the firm's revenue or profit function [e.g., Diewert (1988)] or for the firm's distance function [e.g., Caves, Christensen and Diewert (1982a and 1982b)].

in making these decompositions cancel out in the representation of TFPG as a product of the TP and RS components. However, the decomposition makes us more aware that *an index number TFPG measure typically includes the effects of both technical progress (a shift in the production function) and nonconstant returns to scale if present (a movement along a nonconstant returns to scale production function).*[40]

After defining TP and RS components for the 1–1 case in Subsection 6.1, in Subsection 6.2 theoretical Malmquist output growth, input growth and TFPG indexes are defined for a general multiple input, multiple output production situation.

6.1. Technical progress (TP) and returns to scale (RS) in the simple 1–1 case

The amount of output obtained from the same input volumes could differ in period t versus s for two different sorts of reasons: (1) the same technology might be used, but with a different scale of operation, or (2) the technology might differ. The purpose of the decompositions introduced here is to provide a conceptual framework for thinking about returns to scale versus technological shift changes in TFPG.

In the 1–1 case, TFPG can be measured as the ratio of the period t and s output–input coefficients, as in (2.1-2). Suppose we know the period s and t volumes for the single input and the single output, as well as the true period s and t production functions given by:

$$y_1^s = f^s\left(x_1^s\right) \quad \text{and} \tag{6.1-1}$$

$$y_1^t = f^t\left(x_1^t\right). \tag{6.1-2}$$

Technical progress can be conceptualized as a shift in a production function due to a switch to a new technology for some given scale of operation for the productive process. Four of the possible measures of shift for a production function are considered here. For the first two, the scale is hypothetically held constant by fixing the input level. For the second two, the scale is hypothetically held constant by fixing the output level.

Some hypothetical volumes are needed for defining the four shift measures given here: two on the output side and two on the input side. The output side hypothetical volumes are

$$y_1^{s*} \equiv f^t\left(x_1^s\right) \quad \text{and} \tag{6.1-3}$$

$$y_1^{t*} \equiv f^s\left(x_1^t\right). \tag{6.1-4}$$

The first of these is the output that hypothetically *could be* produced with the scale fixed by the period s input volume x_1^s but using the newer period t technology embodied in f^t. Given technical progress rather than regress, y_1^{s*} should be larger than y_1^s. The second

[40] Favorable or adverse changes in environmental factors facing the firm going from period s to t are regarded as shifts in the production function. We are assuming here that producers are on their production frontier each period; i.e., that they are technically efficient. In a more complete analysis, we could allow for technical inefficiency.

volume, y_1^{t*}, is the output that hypothetically *could be* produced with the scale fixed by the period t input volume x_1^t but using the older period s technology. Given technical progress rather than regress, y_1^{t*} should be smaller than y_1^t.

Turning to the input side now, x_1^{s*} and x_1^{t*} are defined implicitly by

$$y_1^s = f^t\left(x_1^{s*}\right) \quad \text{and} \tag{6.1-5}$$

$$y_1^t = f^s\left(x_1^{t*}\right). \tag{6.1-6}$$

The first of these is the hypothetical amount of the single input factor required to produce the actual period s output, y_1^s, using the more recent period t technology. Given technical progress, x_1^{s*} should be less than x_1^s. The second volume x_1^{t*} is the hypothetical amount of the single input factor required to produce the period t output y_1^t using the older period s technology, so we would usually expect x_1^{t*} to be larger than x_1^t.

The first two of the four technical progress indexes to be defined here are the output based measures given by[41]

$$\text{TP}(1) \equiv y_1^{s*}/y_1^s = f^t\left(x_1^s\right)/f^s\left(x_1^s\right) \quad \text{and} \tag{6.1-7}$$

$$\text{TP}(2) \equiv y_1^t/y_1^{t*} = f^t\left(x_1^t\right)/f^s\left(x_1^t\right). \tag{6.1-8}$$

Each of these describes the percentage increase in output resulting solely from switching from the period s to the period t production technology with the scale of operation fixed by the actual period s or the period t input level for TP(1) and TP(2), respectively.

The other two indexes of technical progress defined here are input based[42]:

$$\text{TP}(3) \equiv x_1^s/x_1^{s*} \quad \text{and} \tag{6.1-9}$$

$$\text{TP}(4) \equiv x_1^{t*}/x_1^t. \tag{6.1-10}$$

Each of these gives the reciprocal of the percentage decrease in input usage resulting solely from switching from the period s to the period t production technology with the scale of operation fixed by the actual period s or the period t output level for TP(3) and TP(4), respectively. That is, for TP(3), technical progress is measured with the output level fixed at y_1^s whereas for TP(4) the output level is fixed at y_1^t.

Each of the technical progress measures defined above is related to TFPG as follows:

$$\text{TFPG} = \text{TP}(i)\,\text{RS}(i) \quad \text{for } i = 1, 2, 3, 4, \tag{6.1-11}$$

where, depending on the selected technical progress measure, the corresponding returns to scale measure is given by

$$\text{RS}(1) \equiv \left[y_1^t/x_1^t\right]/\left[y_1^{s*}/x_1^s\right], \tag{6.1-12}$$

[41] TP(1) and TP(2) are the output based 'productivity' indexes proposed by Caves, Christensen, and Diewert (1982b, p. 1402) for the simplistic case of one input and one output.

[42] TP(3) and TP(4) are the input based 'productivity' indexes proposed by Caves, Christensen, and Diewert (1982b, p. 1407) for the simplistic case of one input and one output.

$$\mathrm{RS}(2) \equiv \left[y_1^{t*}/x_1^t\right]/\left[y_1^s/x_1^s\right], \tag{6.1-13}$$

$$\mathrm{RS}(3) \equiv \left[y_1^t/x_1^t\right]/\left[y_1^s/x_1^{s*}\right], \quad \text{or} \tag{6.1-14}$$

$$\mathrm{RS}(4) \equiv \left[y_1^t/x_1^{t*}\right]/\left[y_1^s/x_1^s\right]. \tag{6.1-15}$$

In the TFPG decompositions given by (6.1-11), the technical progress term, TP(i), can be viewed as a production function *shift*[43] caused by a change in technology, and the returns to scale term, RS(i), can be viewed as a *movement along* a production function with the technology held fixed. Each returns to scale measure will be greater than one if output divided by input increases as we move along the production surface. Obviously, if TP(1) = TP(2) = TP(3) = TP(4) = 1, then RS = TFPG and increases in TFPG are due solely to changes of scale.

For two periods, say $s = 0$ and $t = 1$, and with just one input factor and one output good, the four measures of TP defined in (6.1-7)–(6.1-10) and the four measures of returns to scale defined in (6.1-12)–(6.1-15) can be illustrated graphically, as in Figure 1. (Here the subscript 1 is dropped for both the single input and the single output.)

The lower curved line is the graph of the period 0 production function; i.e., it is the set of points (x, y) such that $x \geqslant 0$ and $y = f^0(x)$. The higher curved line is the graph of the period 1 production function; i.e., it is the set of points (x, y) such that $x \geqslant 0$ and $y = f^1(x)$. The observed data points are A with coordinates (x^0, y^0)

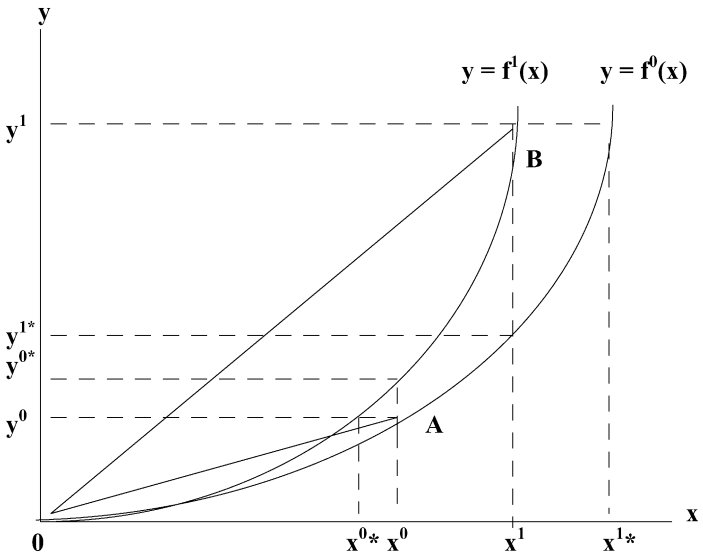

Figure 1. Production function based measures of technical progress.

[43] This shift can be conceptualized as either a move from one production function to another, or equivalently as a change in the location and perhaps the shape of the original production function.

for period 0, and B with coordinates (x^1, y^1) for period 1.[44] Applying formula (2.1-2) from Section 2, for this example we have TFPG $= [y^1/x^1]/[y^0/x^0]$. In Figure 1, this is the slope of the straight line OB divided by the slope of the straight line OA. The reader can use Figure 1 and the definitions provided above to verify that each of the four decompositions of TFPG given by (6.1-11) corresponds to a different combination of shifts of, and movements along, a production function that take us from observed point A to observed point B.[45] Of course, there would be no way of distinguishing among the different possible mechanisms that could yield a move from A to B if nothing were known but the values of the points.

Geometrically, each of the specified measures for the returns to scale is the ratio of two output–input coefficients, say $[y^j/x^j]$ divided by $[y^k/x^k]$ for points (y^j, x^j) and (x^k, y^k) on the *same* fixed production function with $x^j > x^k$. For the ith measure, if the returns to scale component $\mathrm{RS}(i) = [y^j/x^j]/[y^k/x^k]$ is greater than 1, the production function exhibits increasing returns to scale, while if $\mathrm{RS}(i) = 1$ we have constant returns to scale, and if $\mathrm{RS}(i) < 1$ we have decreasing returns. If the returns to scale are constant, then $\mathrm{RS}(i) = 1$ and TP = TFPG.[46] *Note, however, that it is unnecessary to assume constant returns to scale in order to evaluate the index number TFPG measures presented here or in previous sections.*

6.2. *Malmquist indexes*

If the technology for a multiple input, multiple output production process can be represented in each time period by some known production function, this function can be used as a basis for defining theoretical Malmquist volume and Malmquist TFPG indexes. Malmquist indexes are introduced here, and then in the following subsection we show conditions under which these theoretical Malmquist indexes can be evaluated using the same information needed in order to evaluate the TFPG index numbers introduced in Section 3.

Here as previously, we let y_1^t denote the amount of output 1 produced in period t for $t = 0, 1, \ldots, T$. Here we also let $\tilde{y}^t \equiv [y_2^t, y_3^t, \ldots, y_M^t]$ denote the vector of other outputs jointly produced in each period t along with output 1 using the vector of input volumes $x^t \equiv [x_1^t, x_2^t, \ldots, x_N^t]$. Using these notational conventions, the production

[44] In Figure 1, note that if the production function shifts were measured in absolute terms as differences in the direction of the y axis, then these shifts would be given by $y^{0*} - y^0$ (at point A) and $y^1 - y^{1*}$ (at point B). If the shifts were measured in absolute terms as differences in the direction of the x axis, then the shifts would be given by $x^0 - x^{0*}$ (at point A) and $x^{1*} - x^1$ (at point B). An advantage of measuring TP (and TFPG) using ratios is that the relative measures are invariant to changes in the units of measurement whereas the differences are not.

[45] In a regulated industry, increasing returns to scale is often the reason for the regulation. See Diewert (1981b).

[46] Solow's (1957, p. 313) Chart I is similar, but his figure is for the simpler case of constant returns to scale.

functions for output 1 in period s and t can be represented compactly as:

$$y_1^s = f^s(\tilde{y}^s, x^s) \quad \text{and} \quad y_1^t = f^t(\tilde{y}^t, x^t). \tag{6.2-1}$$

Three alternative Malmquist output volume indexes will be defined.[47]

The first Malmquist output index, α^s, is the number which satisfies

$$y_1^t/\alpha^s = f^s(\tilde{y}^t/\alpha^s, x^s). \tag{6.2-2}$$

This index is the number which just deflates the period t vector of outputs, $y^t \equiv [y_1^t, y_2^t, \ldots, y_M^t]$, into an output vector y^t/α^s that can be produced with the period s vector of inputs, x^s, using the period s technology. Due to substitution, when the number of output goods, M, is greater than 1, then the hypothetical output volume vector y^t/α^s will not usually be equal to the actual period s output vector, y^s. However, with only one output good, we have $y_1^t/\alpha^s = f^s(x^s) = y_1^s$ and this Malmquist output index reduces to $\alpha^s = y_1^t/y_1^s$.

A second Malmquist output index, α^t, is defined as the number which satisfies

$$\alpha^t y_1^s = f^t(\alpha^t \tilde{y}^s, x^t). \tag{6.2-3}$$

This index is the number that inflates the period s vector of outputs y^s into $\alpha^t y^s$, an output vector that can be produced with the period t vector of inputs x^t using the period t technology. The index $\alpha^t y^s$ will not usually be equal to y^t when there are multiple outputs. However, when $M = 1$, then $\alpha^t y_1^s = f^t(x^t) = y_1^t$ and $\alpha^t = y_1^t/y_1^s$.

When there is no reason to prefer either the index α^s or α^t, we recommend taking the geometric mean of these indexes. This is the third Malmquist index of output growth, defined as

$$\alpha \equiv [\alpha^s \alpha^t]^{1/2}. \tag{6.2-4}$$

When there are only two output goods, the Malmquist output indexes α^s and α^t can be illustrated as in Figure 2 for $t = 1$ and $s = 0$. The lower curved line represents the set of outputs $\{(y_1, y_2,): y_1 = f^0(y_2, x^0)\}$ that can be produced with period 0 technology and inputs. The higher curved line represents the set of outputs $\{(y_1, y_2,): y_1 = f^1(y_2, x^1)\}$ that can be produced with period 1 technology and inputs. The period 1 output possibilities set will generally be higher than the period 0 one for two reasons: (i) technical progress and (ii) input growth.[48] In Figure 2, the point $\alpha^1 y^0$ is the straight line projection of the period 0 output vector $y^0 = [y_1^0, y_2^0]$ onto the period 1 output possibilities

[47] These indexes correspond to the two output indexes defined in Caves, Christensen, and Diewert (1982b, p. 1400) and referred to by them as Malmquist indexes because Malmquist (1953) proposed indexes similar to these in concept, though his were for the consumer context. Indexes of this sort were subsequently defined as well by Moorsteen (1961) and Hicks (1961, 1981, pp. 192 and 256) for the producer context. See also Balk (1998, Chapter 4).

[48] However, with technical regress, production would become less efficient in period 1 compared to period 0. Also, if the utilization of inputs declined, then the period 1 output production possibilities set could lie below the period 0 one.

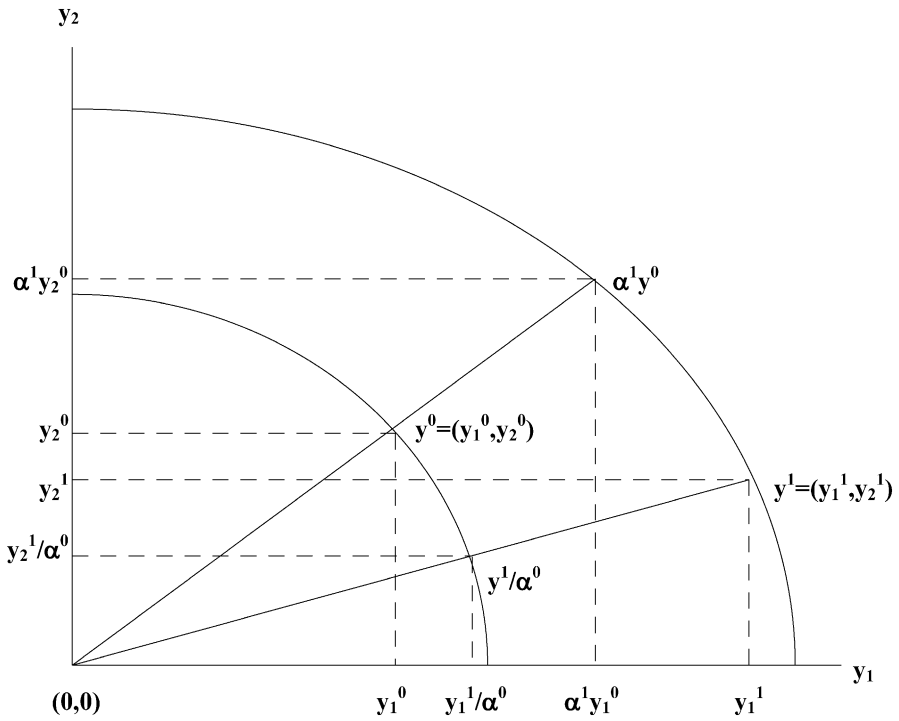

Figure 2. Alternative economic output indexes illustrated.

set, and $y^1/\alpha^0 = [y_1^1/\alpha^0, y_2^1/\alpha^0]$ is the straight line contraction of the output vector $y^1 = [y_1^1, y_2^1]$ onto the period 0 output possibilities set.

We now turn to the input side. A first Malmquist input index, β^s, is defined as follows:

$$y_1^s = f^s\left(\tilde{y}^s, x^t/\beta^s\right) \equiv f^s\left(y_2^s, \ldots, y_M^s, x_1^t/\beta^s, \ldots, x_N^t/\beta^s\right). \tag{6.2-5}$$

This index measures input growth holding fixed the period s technology and output vector. A second Malmquist input index, denoted by β^t, is the solution to the following equation

$$y_1^t = f^t\left(\tilde{y}^t, \beta^t x^s\right) \equiv f^t\left(y_2^t, \ldots, y_M^t, \beta^t x_1^s, \ldots, \beta^t x_N^s\right). \tag{6.2-6}$$

This index measures input growth holding fixed the period t technology and output vector.

When there is no reason to prefer β^s to β^t, we recommend a third Malmquist input index:

$$\beta \equiv \left[\beta^s \beta^t\right]^{1/2}. \tag{6.2-7}$$

Figure 3 illustrates the Malmquist indexes β^s and β^t for the case where there are just two input goods and for $t = 1$ and $s = 0$.

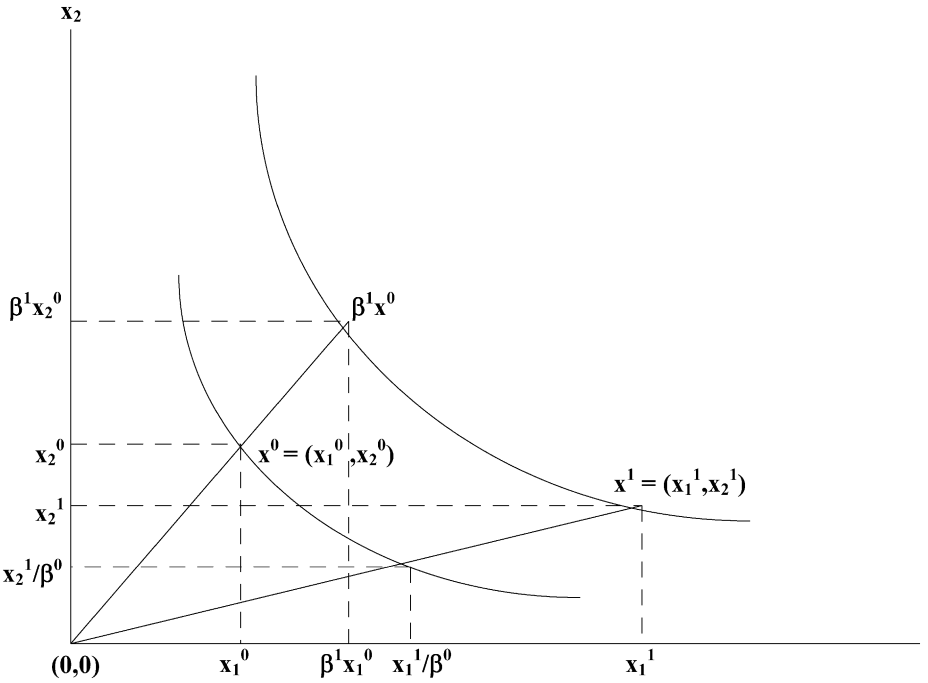

Figure 3. Alternative Malmquist input indexes illustrated.

The lower curved line in Figure 3 represents the set of inputs that are needed to produce the vector of outputs y^0 using period 0 technology. This is the set $\{(x_1, x_2)\colon\ y_1^0 = f^0(\tilde{y}^0, x_1, x_2)\}$. The higher curved line represents the set of inputs that are needed to produce the period 1 vector of outputs y^1 using period 1 technology. This is the set $\{(x_1, x_2)\colon\ y_1^1 = f^1(\tilde{y}^1, x_1, x_2)\}$.[49] The point $\beta^1 x^0 = [\beta^1 x_1^0, \beta^1 x_2^0]$ is the straight line projection of the input vector $x^0 \equiv [x_1^0, x_2^0]$ onto the period 1 input requirements set. The point $x^1/\beta^0 \equiv [x_1^1/\beta^0, x_2^1/\beta^0]$ is the straight line contraction of the input vector $x^1 \equiv [x_1^1, x_2^1]$ onto the period 0 input requirements set.

Once theoretical Malmquist volume indexes have been defined that measure the growth of total output and the growth of total input, then a Malmquist TFPG index for the general N–M case can be defined too. The definition we recommend for the Malmquist TFPG index is

$$\text{TFPG}_M \equiv \alpha/\beta. \tag{6.2-8}$$

[49] If technical progress were sufficiently positive or if output growth between the two periods were sufficiently negative, then the period 1 input requirements set could lie *below* the period 0 input requirements set.

In the 1–1 case, expression (6.2-8) reduces to TFPG(2) as defined in expression (2.1-3), which equals the single measure for TFPG for the 1–1 case.

6.3. *Direct evaluation of Malmquist indexes for the N–M case*

Using the exact index number approach, Caves, Christensen, and Diewert (1982b, pp. 1395–1401) give conditions under which the Malmquist output and input volume indexes $\alpha \equiv [\alpha^s \alpha^t]^{1/2}$ and $\beta \equiv [\beta^s \beta^t]^{1/2}$ defined in (6.2-4) and (6.2-7) equal Törnqvist indexes. More specifically, Caves, Christensen, and Diewert give conditions under which

$$\alpha = Q_T \quad \text{and} \tag{6.3-1}$$

$$\beta = Q_T^*, \tag{6.3-2}$$

where Q_T is the Törnqvist output volume index and Q_T^* is the Törnqvist input volume index. The assumptions required to derive (6.3-1) and (6.3-2) are, roughly speaking: (i) price taking, revenue maximizing behavior, (ii) price taking, cost minimizing behavior, and (iii) a translog technology. Under these assumptions, we can evaluate the theoretical Malmquist measure TFPG_M by taking the ratio of the Törnqvist output and input volume indexes since we have

$$\text{TFPG}_M = \alpha/\beta = Q_T/Q_T^* \equiv \text{TFPG}_T\,. \tag{6.3-3}$$

The practical importance of (6.3-3) is that the Malmquist TFPG index can be evaluated directly from observable prices and volumes without knowing the parameter values for the true period specific production functions. This sort of result can be established as well for other representations of the technology, as we show now.

An intuitive explanation for the remarkable equalities in (6.3-1) and (6.3-2) rests on the following fact: if $f(z)$ is a quadratic function, then we have $f(z^t) - f(z^s) = (1/2)[\nabla f(z^t) + \nabla f(z^s)]^T [z^t - z^s]$. This result follows from applying Diewert's (1976, p. 118) Quadratic Approximation Lemma. Under the assumption of optimizing behavior on the part of the producer, the vectors of first order partial derivatives, $\nabla f(z^t)$ and $\nabla f(z^s)$, will be equal to or proportional to the observed prices. Thus the right-hand side of the above identity becomes observable without econometric estimation.

Recall that the "best" productivity index from the axiomatic point of view is the Fisher productivity index defined in (3.3-4) as

$$\text{TFPG}_F \equiv Q_F/Q_F^*,$$

with the Fisher output volume index Q_F defined by (3.2-3) and input volume index Q_F^* defined by (3.2-6). Diewert (1992b, pp. 240–243) shows these Fisher indexes equal Malmquist indexes when the firm's output distance function over the relevant time span has the functional form

$$\mathrm{d}^t(y,x) = \sigma^t \left[y^T A y \left(x^T C x \right)^{-1} + \alpha^t \cdot y \beta^t \cdot x^{-1} y^T B^t x^{-1} \right]^{1/2}.$$

Here superscript T denotes a transpose, the parameter matrices A and C are symmetric and independent of time t, and the parameter vectors α^t and β^t and also the parameter matrix B^t can depend on time. The vector x^{-1} is defined as consisting of components that are the reciprocals of the components of the vector x of input volumes. The parameter matrices and vectors must also satisfy some additional restrictions that are listed in Diewert (1992b, p. 241).

It should be noted that the above results do *not* rely on the assumption of constant returns to scale in production. *These results extend the concept of superlative index numbers, which were originally defined under the assumption of constant returns to scale*. Also, the assumption of revenue maximizing behavior can be dropped if we know the marginal costs in the two periods under consideration, in which case we could directly evaluate the Malmquist indexes. However, usually we do not know these marginal costs.

In many respects, the Fisher TFPG index is the most attractive index formula.[50] Nevertheless, both the Fisher and the Törnqvist indexes should yield similar results.[51] Both are superlative index numbers. Diewert (1976, 1978b) established that all of the commonly used superlative index number formulas approximate each other to the second order when each index is evaluated at an equal price and volume point.[52] These approximation results, and also Diewert's (1978b) result for the Paasche and Laspeyres indexes, hold *without* the assumption of optimizing behavior and regardless of whether the assumptions about the technology are true. These are findings of numerical rather than economic analysis.

[50] Recall that the Fisher TFPG index satisfies what we have termed the comparability over time ideal, as shown in Subsections 3.3 and 3.4. For an index that satisfies this property, the aggregates that make up the components are comparable for period s and t in the sense that the micro level volumes are aggregated using the same price weights. Diewert (1992b) also shows that the Fisher index satisfies more of the traditional index number axioms than any other formula considered.

[51] See Diewert (1978b, p. 894).

[52] The term superlative means that an index is exact for a flexible functional form. Since the Fisher and the Törnqvist indexes are both superlative, they will both have the same first and second order partial derivatives with respect to all arguments when the derivatives are evaluated at a point where the price and volume vectors take on the same value for both period t and s. T.P. Hill (1993, p. 384) explains current accepted practice as follows: "Thus economic theory suggests that, in general, a symmetric index that assigns equal weight to the two situations being compared is to be preferred to either the Laspeyres or Paasche indices on their own. The precise choice of superlative index – whether Fisher, Törnqvist or other superlative index – may be of only secondary importance as all the symmetric indices are likely to approximate each other, and the underlying theoretic index fairly closely, at least when the index number spread between the Laspeyres and Paasche is not very great". R.J. Hill (2006) showed that whereas the approximation result of Diewert (1978b) which the remarks of T.P. Hill (1993) quoted above are based on and which have found their way into the manuals of statistical agencies around the world do indeed apply to all of the commonly used superlative indexes including the Fisher, Törnqvist, and implicit Törnqvist, the approximation can be poor for some other superlative indexes.

7. Cost function based measures

In this section, we define another set of theoretical output and input growth rate and TFPG measures based on the true underlying cost function instead of the production function as in Section 6. We give conditions under which these indexes equal the Laspeyres and the Paasche indexes. For the two output case, we also show how the Laspeyres and Paasche indexes relate to the Malmquist indexes defined in the previous section.

The period t cost function given by $c^t(y_1, y_2, \ldots, y_M, w_1, w_2, \ldots, w_N)$ in (5-2) is the minimum cost of producing the given volumes $y_1, y_2, \ldots, y_M$ of the M output goods using the input volumes $x_1, x_2, \ldots, x_N$ purchased at the unit prices $w_1, w_2, \ldots, w_N$ and using the period t technology summarized by the production function constraint $y_1 = f^t(y_2, \ldots, y_M, x_1, x_2, \ldots, x_N)$. In this section, we assume that the period s and t cost functions, c^s and c^t, are known and we examine theoretical output, input and productivity indexes that can be defined using these cost functions.

Under the assumptions of perfect information and cost minimizing behavior on the part of the production unit, the actual period t total cost equals the period t cost function evaluated at the period t output volumes and input prices. Thus we have

$$c^t\left(y^t, w^t\right) = \sum_{n=1}^{N} w_n^t x_n^t \equiv w^t \cdot x^t \equiv C^t. \tag{7-1}$$

(As in the above expression, weighted sums will sometimes be represented as inner products of vectors in addition to, or as an alternative to, the summation sign representation.) The cost function in (7-1) is assumed to be differentiable with respect to the components of the vector y at the point (y^t, w^t). Under the assumed conditions, the ith marginal cost for period t, denoted by mc_i^t, is given by

$$\mathrm{mc}_i^t \equiv \partial c^t\left(y^t, w^t\right)/\partial y_i, \quad i = 1, 2, \ldots, M. \tag{7-2}$$

Marginal costs for period s are defined analogously.

Just as the output unit prices were used as weights for the period s and period t volumes in the formulas for the Laspeyres and Paasche volume indexes given in Section 3, here the marginal cost vectors, mc^s and mc^t, are used to define theoretical Laspeyres and Paasche type output and input volume indexes. These indexes are given by

$$\gamma_L \equiv \mathrm{mc}^s \cdot y^t / \mathrm{mc}^s \cdot y^s \quad \text{and} \tag{7-3}$$

$$\gamma_P \equiv \mathrm{mc}^t \cdot y^t / \mathrm{mc}^t \cdot y^s. \tag{7-4}$$

When we have no reason to prefer γ_L over γ_P, we recommend using as a theoretical measure of the output growth rate the geometric mean of γ_L and γ_P; that is, we recommend

$$\gamma \equiv [\gamma_L \gamma_P]^{1/2}. \tag{7-5}$$

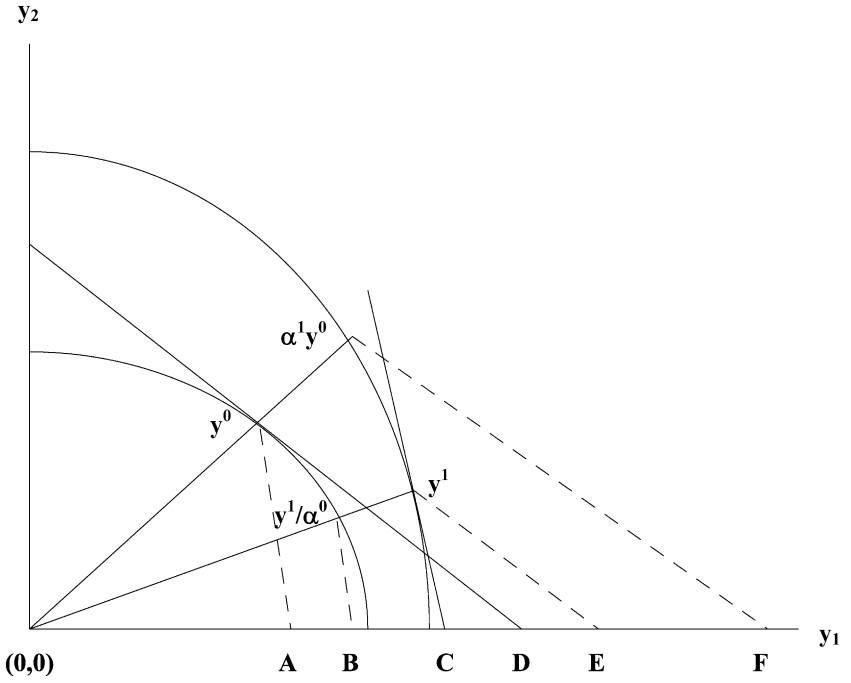

Figure 4. Alternative price based theoretical output indexes.

With price taking, profit maximizing behavior, the observed output volume vector y^t is determined as the solution to the first order necessary conditions for the period t profit maximization problem and economic theory implies that $p^t = \mathrm{mc}^t$. If this is the case, then γ_L defined in (7-3) equals the usual Laspeyres output index, Q_L, defined in (3.2-2), and γ_P defined in (7-4) equals the usual Paasche output index, Q_P, defined in (3.2-1). Moreover, in this case, γ defined in (7-5) equals the Fisher output index, Q_F, defined in (3.2-3).

With just two outputs and under the assumptions of price taking, profit maximizing behavior, the differences between the new theoretical output indexes γ_P and γ_L and the Malmquist output indexes α^0 and α^1 can be illustrated using Figure 4.

The lower curved line in Figure 4 is the period $s = 0$ output possibilities set, $\{(y_1, y_2)\colon\ y_1 = f^0(y_2, x^0)\}$. The higher curved line is the period $t = 1$ output possibilities set, $\{(y_1, y_2)\colon\ y_1 = f^1(y_2, x^1)\}$. The straight line ending in D is tangent to the period 0 output possibilities set at the observed period 0 output vector $y^0 \equiv [y_1^0, y_2^0]$, and the straight line ending in C is tangent to the period 1 output possibilities set at the observed period 1 output vector $y^1 \equiv [y_1^1, y_2^1]$. The marginal costs for period 0 and period 1 are denoted by mc_i^0 and mc_i^1 for outputs $i = 1, 2$. The tangent line through y^0, the output volume vector for period 0, has the slope $-(\mathrm{mc}_1^0/\mathrm{mc}_2^0)$ and the tangent line

through y^1, the period 1 output volume vector, has the slope $-(mc_1^1/mc_2^1)$. The straight line ending in E passes through y^1, and the straight line ending in F passes through $\alpha^1 y^0$. Both of these lines are parallel to the line ending in D, which is the tangent to the period 0 output possibility set at the point (y_1^0, y_2^0). Similarly, the straight line ending in A passes through y^0, and is parallel to the straight line ending in B passes through y^1/α^0, and both are parallel to the line ending in C,[53] which is the tangent to the period 1 output possibility set at the point (y_1^1, y_2^1).

For the theoretical output indexes defined above, we will always have $\gamma_L = \text{OE/OD} < \text{OF/OD} = \alpha^1$ and $\gamma_P = \text{OC/OA} > \text{OC/OB} = \alpha^0$. Although the four output indexes can be quite different in magnitude as illustrated in Figure 4, the geometric average of γ_L and γ_P should be reasonably close to the geometric average of α^0 and α^1. Moving to the input side, the theoretical input volume indexes are given by[54]

$$\delta_L \equiv c^t\left(y^t, w^s\right)/c^s\left(y^s, w^s\right) \quad \text{and} \tag{7-6}$$

$$\delta_P \equiv c^t\left(y^t, w^t\right)/c^s\left(y^s, w^t\right). \tag{7-7}$$

In the case of two inputs and under the assumptions of price taking, profit maximizing behavior, the differences between δ_L and δ_P on the one hand and the Malmquist indexes β^s and β^t on the other hand can be illustrated as in Figure 5. The lower curved line is the period $s = 0$ set of combinations of the two input factors that can be used to produce y^0 under f^0. The upper curved line is the period $t = 1$ set of input combinations that can be used to produce y^1 under f^1.

The straight line ending at the point E in Figure 5 is tangent to the input possibilities curve for period 1 at the observed input vector $x^1 \equiv [x_1^1, x_2^1]$. This tangent line has slope $-(w_1^1/w_2^1)$ and, by construction, the lines ending in A, B, and C have this same slope. The line ending at point C passes through the period 0 observed input vector $x^0 \equiv [x_1^0, x_2^0]$. The line ending at B passes through $x^1/\beta^0 \equiv [x_1^1/\beta^0, x_2^1/\beta^0]$. Finally, the line ending at A is tangent to the period 0 input possibilities set.

Similarly, the straight line ending at the point D in Figure 5 is tangent to the period 0 input possibilities set at the point x^0. The slope of this tangent line is $-(w_1^0/w_2^0)$ and, by construction, the lines ending in F, G, and H have this same slope. The line ending at H passes through x^1. The line ending at G passes through $\beta^1 x^0 \equiv [\beta^1 x_1^0, \beta^1 x_2^0]$, and the line ending at F is tangent to the period 1 input possibilities curve. It can be shown that $\delta_L = \text{OF/OD} < \text{OG/OD} = \beta^1$ and $\delta_P = \text{OE/OA} > \text{OE/OB} = \beta^0$.[55]

[53] Note that the y_1 intercept of a line with the slope of the relevant price ratio – i.e., the y_1 intercept of a line with the slope of the tangent to the designated production possibilities frontier – equals the revenue from the designated output vector denominated in equivalent amounts of good 1.

[54] If there is only one output and if $c^s = c^t$, then δ_L and δ_P reduce to indexes proposed by Allen (1949, p. 199).

[55] The tangency relation follows using Shephard's (1953, p. 11) Lemma: $x_1^0 = \partial c^0(y^0, w_1^0, w_2^0)/\partial w_1$ and $x_2^0 = \partial c^0(y^0, w_1^0, w_2^0)/\partial w_2$. Similarly, the fact that the tangent line ending at E has slope equal to w_1^1/w_2^1

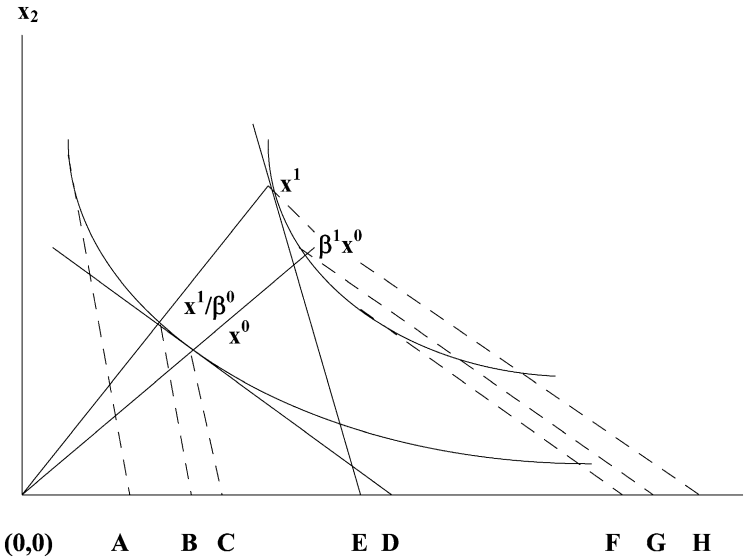

Figure 5. Alternative price based economic input indexes.

8. The Divisia approach

In discrete time approaches to productivity measurement, the price and volume data are defined only for integer values of t, which denotes discrete unit time periods. In contrast, in Divisia's (1926, p. 40) approach, the price and volume variables are defined as functions of continuous time.[56] To emphasize the continuous time feature of the Divisia approach, here the price and volume of output m at time t are denoted by $p_m(t)$ and $y_m(t)$ and the price and volume of input n at time t are denoted by $w_n(t)$ and $x_n(t)$. The price and volume functions are assumed to be differentiable with respect to time over an interval of $0 \leqslant t \leqslant 1$.

Revenue and cost can be represented as

$$R(t) \equiv \sum_{m=1}^{M} p_m(t) y_m(t) \tag{8-1}$$

follows from $x_1^1 = \partial c^1(y^1, w_1^1, w_2^1)/\partial w_1$ and $x_2^1 = \partial c^1(y^1, w_1^1, w_2^1)/\partial w_2$. Note that the x_1 intercept of a line with the slope of $-(w_1^0/w_2^0)$, as is the case for the lines ending in D, F, G or H, or of a line with the slope of $-(w_1^1/w_2^1)$, as is the case for the lines ending in A, B, C or D, is equal to the cost of the stated input vector denominated in units of input factor 1.

[56] For more on the Divisia approach see Hulten (1973) and also Balk (2000).

and

$$C(t) \equiv \sum_{n=1}^{N} w_n(t) x_n(t). \tag{8-2}$$

Differentiating both sides of (8-1) with respect to time and dividing by $R(t)$, we obtain

$$R'(t)/R(t) = \left[\sum_{m=1}^{M} p'_m(t) y_m(t) + \sum_{m=1}^{M} p_m(t) y'_m(t) \right] \Big/ R(t) \tag{8-3}$$

$$= \sum_{m=1}^{M} \left[p'_m(t)/p_m(t) \right] \left[p_m(t) y_m(t)/R(t) \right] + \sum_{m=1}^{M} \left[y'_m(t)/y_m(t) \right] \left[p_m(t) y_m(t)/R(t) \right]$$

$$= \sum_{m=1}^{M} \left[p'_m(t)/p_m(t) \right] s_m^R(t) + \sum_{m=1}^{M} \left[y'_m(t)/y_m(t) \right] s_m^R(t), \tag{8-4}$$

where a prime denotes the time derivative of a function and $s_m^R(t) \equiv [p_m(t) y_m(t)]/R(t)$ is the revenue share of output m at time t. $R'(t)/R(t)$ represents the (percentage) rate of change in revenue at time t.

The first set of terms on the right-hand side of (8-4) is a revenue share weighted sum of the rates of growth in the prices. Divisia (1926, p. 40) defined the aggregate output price growth rate to be[57]

$$P'(t)/P(t) \equiv \sum_{m=1}^{M} \left[p'_m(t)/p_m(t) \right] s_m^R(t). \tag{8-5}$$

The second set of terms on the right-hand side of (8-4) is a revenue share weighted sum of the rates of growth for the volumes of the individual output products. Divisia defined the aggregate output volume growth rate to be

$$Y'(t)/Y(t) \equiv \sum_{m=1}^{M} \left[y'_m(t)/y_m(t) \right] s_m^R(t). \tag{8-6}$$

Substituting (8-5) and (8-6) into (8-4) yields:

$$R'(t)/R(t) = P'(t)/P(t) + Y'(t)/Y(t). \tag{8-7}$$

[57] This is much like declaring the Törnqvist output index to be a measure of output price growth, since it is a weighted aggregate of the growth rates for the prices of the individual output goods.

In words, (8-7) says that the revenue growth at time t is equal to aggregate output price growth plus aggregate output volume growth at time t. Equation (8-7) is the Divisia index counterpart to the output side product test decomposition.

A decomposition similar to (8-7) can be derived in the same way for the (percentage) rate of growth in cost at time t, $C'(t)/C(t)$. Differentiating both sides of (8-2) with respect to t and dividing both sides by $C(t)$ yields

$$\begin{aligned} C'(t)/C(t) &= \left[\sum_{n=1}^{N} w_n'(t)x_n(t) + \sum_{n=1}^{N} w_n(t)x_n'(t)\right] \Big/ C(t) \\ &= \sum_{n=1}^{N} \left[w_n'(t)/w_n(t)\right] s_n^C(t) + \sum_{n=1}^{N} \left[x_n'(t)/x_n(t)\right] s_n^C(t). \end{aligned} \tag{8-8}$$

Here $w_n'(t)$ is the rate of change of the nth input price, $x_n'(t)$ is the rate of change of the nth input volume, and $s_n^C(t) \equiv [w_n(t)x_n(t)]/C(t)$ is the input n share of total cost at time t.

Let $W(t)$ and $X(t)$ denote the Divisia input price and input volume aggregates evaluated at time t, where their proportional rates of change are defined by the two cost share weighted sums of the rates of growth of the individual microeconomic input prices and volumes:

$$W'(t)/W(t) \equiv \sum_{n=1}^{N} \left[w_n'(t)/w_n(t)\right] s_n^C(t) \quad \text{and} \tag{8-9}$$

$$X'(t)/X(t) \equiv \sum_{n=1}^{N} \left[x_n'(t)/x_n(t)\right] s_n^C(t). \tag{8-10}$$

Substituting (8-9) and (8-10) into (8-8) yields the following input side version of (8-7):

$$C'(t)/C(t) = W'(t)/W(t) + X'(t)/X(t). \tag{8-11}$$

In words, (8-11) says that the rate of growth in cost is equal to aggregate input price growth plus aggregate input volume growth at time t. Equation (8-11) is the Divisia index counterpart to the input side product test decomposition in the axiomatic approach to index number theory.

The Divisia TFPG index can be defined as the Divisia measure for the aggregate output volume growth rate, as given in (8-6), minus the Divisia measure for the aggregate input volume growth rate, as given in (8-10)[58]:

$$\text{TFPG}(t) \equiv \left[Y'(t)/Y(t)\right] - \left[X'(t)/X(t)\right], \tag{8-12}$$

[58] See Jorgenson and Griliches (1967, p. 252). Note that the Divisia productivity measure is defined as a difference in rates of growth whereas our previous productivity definitions all involved taking a ratio of growth rates. (Note that the log of a ratio equals the difference of the logs.)

where $Y'(t)/Y(t)$ is given by (8-6) and $X'(t)/X(t)$ is given by (8-10).[59]

A dual expression for TFPG can be derived under the additional assumption that costs equal revenue at each point in time.[60] In this case we have

$$R'(t)/R(t) = C'(t)/C(t), \tag{8-13}$$

and hence the right-hand sides of (8-7) and (8-11) can be equated. Rearranging the resulting equation and applying (8-12) yields:

$$\begin{aligned} \text{TFPG}(t) &\equiv \left[W'(t)/W(t)\right] - \left[P'(t)/P(t)\right] \\ &= \left[Y'(t)/Y(t)\right] - \left[X'(t)/X(t)\right]. \end{aligned} \tag{8-14}$$

Thus, under assumption (8-13), the Divisia TFPG measure equals the Divisia input price growth rate minus the Divisia output price growth rate.

Continuous time formulations can be analytically convenient. Of course, to make them operational for the production of index values, it is necessary to replace derivatives by finite differences. The apparent precision of the Divisia approach vanishes when we do this.[61]

9. Growth accounting

We begin in Section 9 by showing how the growth accounting framework is constructed and its relationship to productivity growth measures and to the exact index number approach. Productivity measures involve comparisons of output and input volume measures, where the volume data are usually derived (as is appropriate) by using price information to transform value data. This same information can be reformulated in a growth accounting framework.

Solow's famous 1957 paper lays out the basics of the growth accounting approach. We take this up in the following Subsection 9.1. We do not attempt to survey the vast growth accounting literature;[62] we seek only to establish the close relationship between growth accounting and productivity measurement for nations.

We complete our brief treatment of growth accounting in Subsection 9.2 with an introduction to the KLEMS approach and the EU KLEMS and World KLEMS initiatives.

[59] For the one output, one input case when $t = 0$, we let $Y(t) = y_1(t) = y(t)$ and $X(t) = x_1(t) = x(t)$. In order to operationalize the continuous time approach, we approximate the derivatives with finite differences as $Y'(0) = y'(0) \cong y(1) - y(0) = y^1 - y^0$ and $X'(0) = x'(0) \cong x(1) - x(0) = x^1 - x^0$. Substituting into (8-12) yields $\text{TFPG}(0) = [y'(0)/y(0)] - [x'(0)/x(0)]$, which is the Divisia approach counterpart to (2.1-3).

[60] See Jorgenson and Griliches (1967, p. 252).

[61] Diewert (1980a, pp. 444–446) shows that there are a wide variety of discrete time approximations to the continuous time Divisia indexes. More recently, Balk (2000) shows how almost any bilateral index number formula can be derived using some discrete approximation to the Divisia continuous time index. Also, as we make the period of time shorter, price and volume data for purchases and sales become "lumpy" and it is necessary to smooth out these lumps. There is no unique way of doing this smoothing.

[62] Virtually all developments in growth accounting are relevant for productivity measurement, and vice versa.

9.1. Solow's 1957 paper

Solow begins with a production function:

$$Y = F(K, L; t), \tag{9.1-1}$$

where Y denotes an output volume aggregate, K and L are aggregate measures for the capital and labor inputs, and t denotes time. A host of index number and aggregation issues are subsumed in the construction of the Y, K and L data series.[63] Solow states that the variable t "for time" appears in the production function F "to allow for technical change". Having introduced t in this way, he goes on to state that this operational definition in no way singles out the adoption of new production technologies. He notes that "slowdowns, speed-ups, improvements in the education of the labor force, and all sorts of things will appear as 'technical change' ".

Solow suggests that we measure technical change by shifts in output associated with the passage of time that are unexplained by increases in expenditures on factor inputs (capital and labor) with all marginal rates of substitution unchanged. This definition of technical change has obvious deficiencies. New technologies are often incorporated into new machinery and new business processes. Solow and others recognized this issue, and a large literature has developed on embodied technical change. However, here we use the original 1957 Solow model because it is a convenient framework for introducing growth accounting, and also for showing how productivity measurement for nations and growth accounting are related. Since Solow assumes that technological change is Hicks neutral in his 1957 paper, the production function in (9.1-1) can be rewritten as

$$Y = A(t) \cdot f(K, L). \tag{9.1-2}$$

That is, the production function can be decomposed into a time varying multiplicative technical change term and an atemporal production function.[64] The multiplicative factor, $A(t)$, represents the effects of shifts over time after controlling for the growth of K and L.

Solow's 1957 study represented a reconciliation of the forecasting results for early estimated aggregate production functions with direct measures of the growth of aggregate product. Abramovitz (1956) had previously compared a weighted sum of labor and capital inputs with a measure of total output and had concluded that to reconcile these,

[63] Some studies such as Hall (1990) essentially treat the economy of a nation as though it produced the single output of income or GDP. However, from one time period to another (or one nation to another) the product mix that makes up national output can shift. See Diewert, Nakajima, A. Nakamura, E. Nakamura and M. Nakamura (2007) – DN4 for short.

[64] Solow's recommendations in his 1957 paper encouraged other researchers to be interested in measuring efficiency improvement in their econometric studies by the ratio of period t and period s efficiency parameters, with the production function for each period specified as the product of a time varying efficiency parameter and an atemporal production function f.

it was necessary to invoke a positive role for technical progress over time. He recommended using time itself as a proxy for productivity improvements. Still earlier, in a 1942 German article, Tinbergen made use of an aggregate production function that incorporated a time trend. His stated purpose in doing this was to capture changes over time in productive efficiency.

In his 1957 paper, Solow re-formulates the output and capital input variables as $(Y/L) = y$ and $(K/L) = k$. Notice that y is output per unit of labor input: a labor productivity index.[65] He specifies that the production function is homogeneous of degree one (thereby assuming constant returns to scale), and that capital and labor are paid their marginal products so that total revenue equals the sum of all factor costs.

Making use of the Divisia methodology, Solow arrives at the following growth accounting equation[66]:

$$\dot{y}/y = (\dot{A}/A) + s_K(\dot{k}/k), \tag{9.1-3}$$

where the dots over variables denote time derivatives, and s_K stands for the national income share of capital.[67] Solow approximates the term $(\dot{A}/A)$ in (9.1-3) by $(\Delta A/A)$. He uses similar discrete approximations for the other variables, and rearranges terms to obtain

$$(\Delta A/A) = (\Delta y/y) - s_k(\Delta k/k). \tag{9.1-4}$$

Solow then produces values for $A(t)$ for the years of 1910 through 1949 by setting $A(1909) = 1$ and using the formula $A(t+1) = A(t)[1 + \Delta A(t)/A(t)]$.

Solow computes his productivity growth values – the values for $(\Delta A/A)$ – using index number rather than econometric methods. The correspondence he establishes between the functional form he assumes for the production function and this productivity growth measure is an application of the exact approach to index numbers (outlined in Section 5).

The growth accounting literature grew phenomenally from 1957 on. The methodology was extended and applied in large scale empirical studies by Griliches (1960, 1963), Denison (1967) and Kendrick (1973, 1976, 1977) and by Dale W. Jorgenson and his colleagues. In his Presidential Address delivered at the one-hundred tenth meeting of the American Economic Association, Harberger (1998, p. 1) describes growth accounting as an important success story for the economics profession, and asserts that the work of Jorgenson and Griliches (1967), Jorgenson, Gollop and Fraumeni (1987), and Jorgenson (1995a, 1995b) has carried growth accounting to the level of a "high art".

[65] If L is measured as an aggregate of hours for different types of labor weighted by their respective average wages, then this is a wage weighted hours labor productivity measure, as defined in (2.3-4). If L is total (unweighted) hours of work, then y is hours labor productivity, as defined in (2.3-5) in Section 2, whereas if L is measured as the number of workers, then y is worker based labor productivity defined in (2.3-6).

[66] The Divisia productivity index, defined by (8-12) in Section 8 of our paper, was related to measures of production function shift by Solow (1957) for the two input, one output case, and by Jorgenson and Griliches (1967) for the general N input, M output case.

[67] Solow assumes that all factor inputs can be classified as capital or labor; hence $s_L = 1 - s_K$ is the national income share of labor.

9.2. Intermediate goods and the KLEMS approach

We come now to the question of how intermediate goods should be treated. Not all current period production in a nation is for final demand. Many firms sell some or all of their output to other firms as intermediate inputs. For example, increasing numbers of firms are outsourcing business services such as call center and accounting operations. Some of the outsourcing takes place with other firms in the same nation, but increasing amounts are with firms in other nations (the so-called "off shoring").

Output can be measured as value added, or as gross output. GNP and GDP are both value added measures, despite the fact that these terms begin with the word "gross". GNP and GDP are value added measures because they exclude intermediate inputs (i.e., they exclude produced and purchased energy and goods and services used in the production of final demand products). In contrast, a gross output measure includes the intermediate products. Either a value added or a gross output measure can be used in a growth accounting study and in specifying any of the productivity measures that have been discussed in previous sections, but the results will differ depending on this choice.

The difference between the two output concepts is less pronounced at the national level than it is at the sectoral or industry level. At the aggregate level, gross output and value added measures differ only to the extent that intermediate inputs are part of international trade.[68]

However, for the economy of a nation as a whole, changes in intermediate input usage can have productivity impacts (using either a gross or value added output measure). Research efforts to understand productivity impacts with their origin in intermediate product usage will be hampered if we do not have data on these inputs. For example, there can still be ongoing substitution effects between factor inputs such as labor and intermediate inputs, especially including business services through outsourcing and off shoring.[69] Also, modern productivity improvement techniques are aimed at improving the efficiency with which both intermediate and primary inputs are used. For example, in the manufacturing sector, just-in-time production, statistical process control, computer-aided design and manufacturing, and other such processes reduce error rates and cut down on sub-standard rejected production. In so doing, they reduce the wastage of materials as well as workers' time. Such efficiencies should probably be taken into account in measuring productivity growth.

An advantage of gross output measures is that they acknowledge and allow for intermediate inputs as a source of industry growth. In this sense, they provide a more complete picture of the production process [Sichel (2001, p. 7)]. It is true that the net productivity measures based on value added reflect savings in intermediate inputs

[68] At the industry or sector level, intermediate usage tends to be a much higher proportion of gross output. See Hulten (1978).

[69] This is demonstrated, for instance, by Gullickson and Harper (1999). Price and output measurement in many areas of business services are problematical, including core banking services. See Wang and Basu (2007).

because real value added per unit of primary input rises when unit requirements for intermediate inputs are reduced, but the effect is not explicit. Gross output-based measures explicitly indicate the contribution of savings in intermediate inputs. The deflation of gross output is conceptually straightforward too. An index of the nominal value of output is divided by an output price index to derive a volume index of gross output.

The deflation of value added output is complex. It involves double deflation because the volume change for value added combines the volume change of gross output and intermediate inputs. The term 'double' indicates that both production and intermediate inputs must be deflated in order to measure changes in the real output attributable to the factors of production in an industry.

Since value added is defined as the difference between separately deflated gross output and intermediate inputs, the use of value added as a measure of output in productivity studies imposes restrictions on the generality of the model of producer behaviour and on the role of technological change [see Diewert (1980b)]. The implied model of sectoral production does not allow for substitution possibilities between the elements of the value added function (capital and labor) and intermediate inputs. For example, it assumes that price changes in intermediate inputs do not influence the relative use of capital and labor. It restricts the role of technological change by assuming that such change only affects the usage of capital and labor.

With appropriate treatment of intermediate inputs, a mutually consistent set of estimates can be obtained at each level of economic activity. This is one objective of the KLEMS (capital, labor, energy, materials and services) approach. This approach is important because consistent aggregation is necessary to answer questions about the contribution of individual industries to overall national economic growth and productivity growth.

Jorgenson, Gollop and Fraumeni (1987) were the first scholars to work out and apply the basic KLEM methodology for a detailed industry analysis of productivity growth in the post-war US economy.[70]

The primary aim of the European KLEMS (EU KLEMS) project is to arrive at an internationally comparable dataset for a KLEMS-type analysis of productivity growth for European countries. Originally there were eight participating nations – Denmark, Finland, France, Germany, Italy, Netherlands, Spain and the United Kingdom – but the list soon grew to more than 30.[71] The World KLEM project, of which EU KLEM is the first component, represents an international platform for national level research and data collection efforts with a clear emphasis on the need for international comparability.

[70] For more on the development of the KLEMS approach in the United States, see Dean and Harper (2000), Gullickson (1995), and Gullickson and Harper (1999) and also Jorgenson (2001), Gollop (1979), and Gollop and Jorgenson (1980, 1983).

[71] In addition, the dataset, which includes the development of purchasing power parities, can be used for other purposes such as the analysis of international competitiveness and investment opportunities. It can serve as a base for further research into for example the impact of high-tech industries or human capital building on economic growth and productivity change. For more information, and for free use of the EU KLEMS database go to http://www.euklems.org/.

All of the productivity measures introduced in this paper can be recast in a KLEMS formulation. TFP or MFP growth as measured by the value added method will systematically exceed the index values based on gross output by a factor equal to the ratio of gross output to value added.[72] Productivity in the gross output formulation is $Y/(E + M + L + K)$ where Y is gross output, E is energy, M is materials, L is labor input and K is capital input. Productivity in the real value added framework is roughly $(Y - E - M)/(L + K)$. Given a productivity improvement of ΔY with all inputs remaining constant, the gross output productivity growth rate is

$$\begin{aligned}&\big((Y + \Delta Y)/(K + L + E + M)\big)\big/\big(Y/(K + L + E + M)\big)\\&\quad = (Y + \Delta Y)/Y = 1 + (\Delta Y/Y),\end{aligned} \tag{9.2-1}$$

which is less than the real value added productivity growth rate of

$$\begin{aligned}&\big((Y + \Delta Y - E - M)/(K + L)\big)\big/\big((Y - E - M)/(K + L)\big)\\&\quad = 1 + \big(\Delta Y/(Y - E - M)\big).\end{aligned} \tag{9.2-2}$$

Thus, the smaller denominator in the value added productivity measure translates into a larger productivity growth measure.[73] Several studies have found that productivity growth measured according to a value added model is greater than that derived from a model that also takes intermediate inputs into account.[74]

Diewert (2002a) notes that industry estimates of output and intermediate input are fragile in all countries due to the lack of adequate surveys on *intermediate input flows* and in particular, of *service flows* between industries.

10. Improving the model

The basic framework for productivity measurement and growth accounting for nations continues to be improved. Here we consider two of the areas of development: the specification of the measure of national output (Subsection 10.1), and efforts to relax the assumption of constant returns to scale that has been a central feature of the conventional productivity measurement and growth accounting framework (Subsection 10.2).

[72] See Diewert (2002a, p. 46, endnote 21).

[73] See also Schreyer (2001, p. 26).

[74] For example, Oulton and O'Mahony (1994) show that the value added method produces estimates of MFP growth for manufacturing in the United Kingdom that are roughly twice those given by the gross output method. It is to be expected, of course, the sub-national level studies will be more affected by the choice of a value added or gross output measure. For example, van der Wiel (1999) shows that MFP estimates for various Dutch industries are much larger for the value added than for the gross output method.

10.1. Different concepts of national product and income

Economists have long argued that net domestic product (NDP) is the proper measure of national output for welfare analyses.[75] Yet most studies of the economic strength of a country use gross domestic product (or sometimes gross national product, GNP, as in Solow's 1957 paper) as "the" measure of output, as we did too in the previous sections of this paper. The difficulty of devising satisfactory measures of depreciation is a key reason for the dominance of the GDP and GNP measures.[76] However, by deducting even a very imperfect measure of depreciation (and obsolescence) from gross investment, we could probably come closer to a measure of output that could be consumed period after period without impairing future production possibilities.[77]

Each definition of net product gives rise to a corresponding definition of "income". In the economics literature, most of the discussion of alternative measures of net output has been conducted in terms of alternative "income" measures, so here we follow the literature and discuss alternative "income" measures rather than alternative measures of "net product". The key ideas can be understood by considering alternative income concepts in a very simple two period ($t = 0, 1$) economy with only two goods: consumption C^t with unit price p_C^t and a durable capital input K^t. Net investment I^t during period t is defined as the end of the period capital stock, K^t, less the beginning of the period capital stock, K^{t-1}: i.e., $I^t \equiv K^t - K^{t-1}$.

Samuelson (1961, p. 45) used the Marshall (1890)–Haig (1921/1959) definition of income as consumption plus the consumption equivalent of the increase in net wealth over the period, and we follow his example in this regard. Nominal income in period 1 can be represented as $p_C^1 C^1 + p_I^1 I^1$ where I^1 can be defined as net investment in period 1.

Net investment can be redefined in terms of the difference between the beginning and end of period 1 capital stocks. If we substitute this representation of net investment into Samuelson's definition of period 1 nominal income, we obtain the following definition for *period 1 nominal income*:

$$\begin{aligned} \text{Income } A &\equiv p_C^1 C^1 + p_I^1 I^1 = p_C^1 C^1 + p_I^1 \left(K^1 - K^0\right) \\ &= p_C^1 C^1 + p_I^1 K^1 - p_I^1 K^0. \end{aligned} \tag{10.1-1}$$

Here, the beginning and end of period capital stocks are valued at the same price, p_I^1.

On conceptual grounds, it might be more reasonable to value the beginning of the period capital stock at the beginning of the period opportunity cost of capital, p_K^0, and

[75] For a closed economy, there is no distinction between net domestic product (NDP) and net national product (NNP), but the economies of countries like the United States, Canada and Japan are not closed, and the term globalization that is often used in conjunction with commentaries on the way the world economic situation is changing describes a condition of increasing openness.

[76] On the treatment of depreciation effects in the US statistics, see Fraumeni (1997). See also Hulten and Wykoff (1981a, 1981b). For a more current and international perspective and references, see T.P. Hill (2005). This topic has long occupied economists. See, for example, Hotelling (1925).

[77] This material is developed more fully in Diewert (2006d) and Diewert and Schreyer (2006b).

the end of the period capital stock at the end of the period expected opportunity cost of capital, p_K^1. That is, perhaps we should replace p_I^1 in (10.1-1) by p_K^1 for the K^1 portion of $I^1 = K^1 - K^0$, and by p_K^0, adjusted for the effects of inflation over the duration of period 1, for the K^0 portion.[78] To adjust p_K^0 for inflation we could use either a capital *specific price index*, denoted here by $1 + i^0$, or a *general price index* that is based on the movement of consumer prices, denoted by $1 + \rho^0$:

$$1 + i^0 \equiv p_K^1 / p_K^0 \quad \text{or} \tag{10.1-2}$$

$$1 + \rho^0 \equiv p_C^1 / p_C^0. \tag{10.1-3}$$

These alternative adjustment factors lead to different *measures of income from the perspective of the level of prices prevailing at the end of period 1*:

$$\text{Income B} \equiv p_C^1 C^1 + p_K^1 K^1 - \left(1 + i^0\right) p_K^0 K^0, \tag{10.1-4}$$

$$\text{Income C} \equiv p_C^1 C^1 + p_K^1 K^1 - \left(1 + \rho^0\right) p_K^0 K^0. \tag{10.1-5}$$

Comparing (10.1-4) and (10.1-1), it is easily seen that Income B equals Income A. Thus, for a measure of output, we are left with the options of choosing between Income A, which is adjusted for (i.e., net of) wear and tear,[79] and Income C, which is adjusted for wear and tear and also anticipated revaluation,[80] or of sticking with a gross output measure.

The "traditional" user cost of capital (which approximates a market rental rate for the services of a capital input for the accounting period), u^1, consists of three additive terms:

$$u^1 = U^1 + D^1 + R^1, \tag{10.1-6}$$

where U^1 denotes the reward for waiting (an interest rate term), D^1 denotes the cross sectional depreciation term (the wear and tear depreciation term), and R^1 is the anticipated revaluation term which can be interpreted as an obsolescence charge if the asset is anticipated to fall in price over the accounting period. The gross output income concept corresponds to the traditional user cost term u^1. This gross income measure can be used as an approximate indicator of short run production potential, but it is not suitable for use as an indicator of sustainable consumption. For an indicator of sustainable consumption, income concept A or C is more appropriate.

Expressed in words, for Income A, we take the wear and tear component of the traditional user cost, D^1, times the beginning of period corresponding capital stock, K^0, out of the primary input category and treat this as a negative offset to the period's gross investment. Diewert (2006d) suggests that the Income A

[78] In order to simplify our algebra, we will assume that it is not necessary to adjust p_C^1 into an end of period 1 price.

[79] We can associate this income concept with Marshall (1890), Haig (1921/1959), Pigou (1941) and Samuelson (1961). On machine replacement issues, see, for example, Cooper and Haltiwanger (1993).

[80] We can associate this income concept with Hayek (1941), Sterling (1975) and T.P. Hill (2000).

concept can be interpreted as a *maintenance of physical capital approach* to income measurement. In terms of the Austrian production model favored by Hicks (1939, 1940, 1942, 1946, 1961, 1973) and by Edwards and Bell (1961), capital at the beginning and end of the period ($K^0$and K^1, respectively) should both be valued at the end of period stock price for a unit of capital, p_K^1, and the contribution of capital accumulation to current period income is simply the difference between the end of period value of the capital stock and the beginning of the period value (at end of period prices), $p_K^1 K^1 - p_K^1 K^0$. This difference between end and beginning of period values for the capital stock can be converted into consumption equivalents and then can be added to actual period 1 consumption in order to obtain Income A.

Income C can be computed by subtracting from gross output both wear and tear depreciation, $D^1 K^0$, and the revaluation term, $R^1 K^0$, and treating both of these terms as negative offsets to the period's gross investment.[81] Diewert (2006d) terms this a *maintenance of real financial capital* approach to income measurement.

In the Austrian production model tradition followed by Hicks (1961) and Edwards and Bell (1961), capital stocks at the beginning and end of the period should be valued at the prices prevailing at the beginning and the end of the period,[82] p_K^0 and p_K^1 respectively, and then these beginning and end of period values of the capital stock should be converted into consumption equivalents (at the prices prevailing at the beginning and end of the period). Thus the end of the period value of the capital stock is $p_K^1 K^1$ and this value can be converted into consumption equivalents at the consumption prices prevailing at the end of the period. The beginning of the period value of the capital stock is $p_K^0 K^0$. To convert this value into consumption equivalents at end of period prices, we must multiply this value by $(1 + \rho^0)$, which is one plus the rate of consumer price inflation over the period. This price level adjusted difference between end and beginning of period values for the capital stock, $p_K^1 K^1 - (1 + \rho^0) p_K^0 K^0$, can be converted into consumption equivalents and then can be added to actual period 1 consumption in order to obtain Income C.

The difference between Income A and Income C can be viewed as follows. Income A (asymmetrically) uses the end of period stock price of capital to value both the beginning and end of period capital stocks and then converts the resulting difference in values into consumption equivalents at the prices prevailing at the end of the period. In contrast, Income C symmetrically values beginning and end of period capital stocks at the stock prices prevailing at the beginning and end of the period and *directly* converts these values into consumption equivalents and then adds the difference in these consumption equivalents to actual consumption.

[81] The resulting Income 3 can be interpreted to be consistent with the position of Hayek (1941), Sterling (1975) and T.P. Hill (2000).

[82] Strictly speaking, the end of period price is an expected end of period price.

In symbols, the difference between income concepts A and C is as follows:

$$\begin{aligned}&\text{Income A} - \text{Income C}\\&\quad = p_C^1 C^1 + p_I^1 K^1 - p_I^1 K^0 - \left[p_C^1 C^1 + p_K^1 K^1 - \left(1+\rho^0\right) p_K^0 K^0\right]\\&\quad = \left(\rho^0 - i^0\right) p_K^0 K^0. \qquad (10.1\text{-}7)\end{aligned}$$

If ρ^0 (the general consumer price inflation rate) is greater than i^0 (the asset inflation rate) over the course of the period, then there is a negative real revaluation effect (so that obsolescence effects dominate). In this case, Income C will be less than Income A, reflecting the fact that capital stocks have become less valuable (in terms of consumption equivalents) over the course of the period. If ρ^0 is less than i^0 over the course of the period, then the real revaluation effect is positive (so that capital stocks have become more valuable over the period). In this case, Income C exceeds Income A, reflecting the fact that capital stocks have become more valuable over the course of the period and this real increase in value contributes to an increase in the period's income which is not reflected in Income A.

Both Income A and Income C have reasonable justifications. Choosing between them is not a straightforward matter. Income A is easier to justify to national income accountants because it relies on the standard production function model. However, we lean towards Income C over Income A for three reasons: (i) It seems to us that (expected) obsolescence charges are entirely similar to normal depreciation charges and Income C reflects this similarity. (ii) In contrast to Income A, Income C does not value the beginning and end of period value of the capital stock in an asymmetric manner. And (iii) it seems to us that waiting services ($U^1 K^0$) along with labor services and land rents are natural primary inputs whereas depreciation and revaluation services ($D^1 K^0$ and $R^1 K^0$, respectively) are more naturally regarded as intermediate input charges.[83]

10.2. Relaxing the constant returns to scale assumption

There has also been strong and persistent interest in finding theoretically palatable and empirically feasible ways to relax the assumption of constant returns to scale in the growth accounting and productivity measurement literatures. Denny, Fuss and Waverman (1981, pp. 196–199) relate the Divisia TFP measure, given in Section 8 by (8-12), to shifts in the cost function without making the assumption of constant returns to scale. Here we summarize the analysis of Denny, Fuss and Waverman using slightly different notation than they did.

Our discussion of Divisia indexes in Section 8 made no mention of cost minimizing behavior. In contrast, the approach of Denny, Fuss and Waverman requires us to assume

[83] Income C is based on the Austrian model of production which has its roots in the work of Böhm-Bawerk (1891), von Neumann (1937) and Malinvaud (1953) but these authors did not develop the user cost implications of the model. On the user cost implications of the Austrian model, see Hicks (1973, pp. 27–35) and Diewert (1977, pp. 108–111, 1980a, pp. 472–474).

that the productive unit continuously minimizes costs over the time period of interest: $0 \leqslant t \leqslant 1$. The production unit's cost function will be written here as $c(y, w, t)$ to emphasize the treatment of time as continuous, where $y(t) \equiv [y_1(t), \ldots, y_M(t)]$ denotes the vector of outputs and $w(t) \equiv [w_1(t), \ldots, w_N(t)]$ denotes the vector of input prices.[84] (The t variable in $c(y, w, t)$ is viewed as representing the fact that the cost function is continuously changing due to technical progress.) Under the assumption of cost minimizing behavior, for $0 \leqslant t \leqslant 1$, we have

$$C(t) \equiv \sum_{n=1}^{N} w_n(t)x_n(t) = c\big[y(t), w(t), t\big]. \tag{10.2-1}$$

We define the continuous time technical progress measure as minus the (percentage) rate of increase in cost at time t:

$$\mathrm{TP}(t) \equiv -\big\{\partial c\big[y(t), w(t), t\big]/\partial t\big\}/c\big[y(t), w(t), t\big]. \tag{10.2-2}$$

Shephard's (1953, p. 11) Lemma implies that the partial derivative of the cost function with respect to the nth input price equals the cost minimizing demand for input n, given by

$$x_n(t) = \partial c\big[y(t), w(t), t\big]/\partial w_n, \quad n = 1, 2, \ldots, N. \tag{10.2-3}$$

Differentiating both sides of (10.2-1) with respect to t, dividing both sides of the resulting equation by $C(t)$, and using (10.2-2) and (10.2-3), we obtain

$$\begin{aligned} C'(t)/C(t) &\equiv \sum_{m=1}^{M} \big\{\partial c\big[y(t), w(t), t\big]/\partial y_m\big\}\big[y'_m(t)/C(t)\big] \\ &\quad + \sum_{n=1}^{N} x_n(t)\big[w'_n(t)/C(t)\big] - \mathrm{TP}(t) \\ &= \sum_{m=1}^{M} \varepsilon_m(t)\big[y'_m(t)/y_m(t)\big] + \sum_{n=1}^{N} s_n^C(t)\big[w'_n(t)/w_n(t)\big] - \mathrm{TP}(t), \end{aligned} \tag{10.2-4}$$

where

$$\varepsilon_m(t) \equiv \big\{\partial c\big[y(t), w(t), t\big]/\partial y_m\big\}\big/\big\{c\big[y(t), w(t), t\big]/y_m(t)\big\}$$

is the elasticity of cost with respect to the mth output volume and

$$s_n^C(t) \equiv \big[w_n(t)x_n(t)\big]/C(t)$$

is the nth input cost share.

[84] To reconcile the notation used here with the notation used in Sections 2–8, note that

$$c^0\big(y^0, w^0\big) = c\big[y(0), w(0), 0\big] \quad \text{and} \quad c^1\big(y^1, w^1\big) = c\big[y(1), w(1), 1\big]$$

with $y(t) \equiv y^t$ and $w(t) \equiv w^t$ for $t = 0, 1$.

Denny, Fuss and Waverman (1981, p. 196) define the rate of change of the continuous time output aggregate, $Q(t)$, as follows:

$$Q'(t)/Q(t) \equiv \sum_{m=1}^{M} \varepsilon_m(t)\big[y'_m(t)/y_m(t)\big] \Big/ \sum_{i=1}^{M} \varepsilon_i(t). \tag{10.2-5}$$

Recall that the Divisia expression for the output growth rate given in (8-6) weights the individual output growth rates, $y'_m(t)/y_m(t)$, by the revenue shares, $s_m^R(t)$. Alternatively, in (10.2-5), $y'_m(t)/y_m(t)$ is weighted by the mth cost elasticity share, $\varepsilon_m(t)/\sum_{i=1}^{M} \varepsilon_i(t)$. It can be shown that $\sum_{i=1}^{M} \varepsilon_i(t)$ is the percentage increase in cost due to a one percent increase in scale for each output.[85] We define the reciprocal of this sum to be a measure of (local) returns to scale:

$$\mathrm{RS}(t) \equiv \left[\sum_{i=1}^{M} \varepsilon_i(t)\right]^{-1}. \tag{10.2-6}$$

Now equate the right-hand side of (8-11) to the right-hand side of (10.2-4). Using (8-9), (10.2-5), and (10.2-6), we obtain the following decomposition of the technical progress measure in terms of returns to scale, output growth and input growth:

$$\mathrm{TP}(t) = \big[\mathrm{RS}(t)\big]^{-1}\big[Q'(t)/Q(t)\big] - \big[X'(t)/X(t)\big]. \tag{10.2-7}$$

In order to relate the technical progress measure $\mathrm{TP}(t)$ defined by (10.2-7) to the Divisia productivity measure $\mathrm{TFPG}(t)$ defined by (8-12), we use Equation (8-12) to solve for $X'(t)/X(t) = [Y'(t)/Y(t)] - \mathrm{TFPG}(t)$ and then solve for $X'(t)/X(t)$. Equating these two expressions for $X'(t)/X(t)$ and rearranging terms yields

$$\mathrm{TFPG}(t) = \big[Y'(t)/Y(t)\big] - \big[\mathrm{RS}(t)\big]^{-1}\big[Q'(t)/Q(t)\big] + \mathrm{TP}(t) \tag{10.2-8}$$

$$= \mathrm{TP}(t) + \big\{Q'(t)/Q(t)\big\}\big\{1 - \big[\mathrm{RS}(t)\big]^{-1}\big\} + \big\{\big[Y'(t)/Y(t)\big] - \big[Q'(t)/Q(t)\big]\big\}. \tag{10.2-9}$$

Equation (10.2-8) is due to Denny, Fuss, and Waverman (1981, p. 197). This equation says that the Divisia productivity index equals the technical progress measure $\mathrm{TP}(t)$ plus

[85] The elasticity of cost with respect to a scale variable k is defined as $\{1/c[y(t), w(t), t]\}$ times the following derivative evaluated at $k = 1$:

$$\partial c\big[ky(t), w(t), t\big]/\partial k = \sum_{m=1}^{M} y_m(t)\partial c\big(y(t), w(t), t\big)/\partial y_m = c\big[y(t), w(t), t\big] \sum_{m=1}^{M} \varepsilon_m(t),$$

where the last equality follows from the definition of $\varepsilon_m(t)$ below (10.2-4). Therefore, the elasticity of cost with respect to scale equals

$$\big\{1/c\big[y(t), w(t), t\big]\big\}\big\{c\big[y(t), w(t), y\big]\big\} \sum_{m=1}^{M} \varepsilon_m(t) = \sum_{m=1}^{M} \varepsilon_m(t).$$

the marginal cost weighted output growth index, $Q'(t)/Q(t)$, times a term that depends on the returns to scale term, $\{1 - [\mathrm{RS}(t)]^{-1}\}$, and that will be positive if and only if the local returns to scale measure $\mathrm{RS}(t)$ is greater than 1, plus the difference between the Divisia output growth index, $Y'(t)/Y(t)$, and the marginal cost weighted output growth index, $Q'(t)/Q(t)$.

Denny, Fuss, and Waverman (1981, p. 197) interpret the term $Y'(t)/Y(t) - Q'(t)/Q(t)$ as the effect on TFPG of nonmarginal cost pricing of a nonproportional variety. Their argument goes like this. Suppose that the mth marginal cost is proportional to the period t selling price $p_m(t)$ for $m = 1, 2, \ldots, M$. Let the common factor of proportionality be $\lambda(t)$. Then we have:

$$\partial c\big[y(t), w(t), t\big]/\partial y_m = \lambda(t) p_m(t), \quad m = 1, 2, \ldots, M. \tag{10.2-10}$$

Using (10.2-10) together with the definitions of $\varepsilon_m(t)$ and $s_m^R(t)$, we find that

$$\varepsilon_m(t) = s_m^R(t)\lambda(t)R(t)/C(t), \quad m = 1, 2, \ldots, M. \tag{10.2-11}$$

Substituting (10.2-11) into (10.2-4) and using (8-6) yields

$$Y'(t)/Y(t) = Q'(t)/Q(t). \tag{10.2-12}$$

If marginal costs are proportional to output prices[86] so that (10.2-10) holds, then the term $Y'(t)/Y(t) - Q'(t)/Q(t)$ vanishes from (10.2-9).[87] This approach provides a continuous time counterpart to the economic approaches to productivity measurement developed in previous sections.

Since the 1981 Denny–Fuss–Waverman paper was published, many others have worked on finding empirically tractable ways of treating nonconstant returns to scale in growth accounting and productivity analysis, and on dealing with the associated issue of imperfect markets and markups.

The traditional approach to estimating returns to scale is to define the elasticity of scale in the context of a producer behavioral relationship, and then estimate that parameter along with all the others for the behavioral relationship. This approach tends to be plagued by degrees of freedom and multicollinearity problems. Building on the original results of Yoshioka, Nakajima and M. Nakamura (1994) in a 2007 paper, Diewert, Nakajima, A. Nakamura, E. Nakamura and M. Nakamura (DN4 for short) extend

[86] It can be shown that if the firm (i) maximizes revenues holding constant its utilization of inputs and (ii) minimizes costs holding constant its production of outputs, then marginal costs will be proportional to output prices; i.e., we obtain $p^t/p^t \cdot y^t = \mathrm{mc}^t/\mathrm{mc}^t \cdot y^t$. Hence prices in period t, p^t, are proportional to marginal costs, mc^t. Note that assumptions (i) and (ii) above are weaker than the assumption of overall profit maximizing behavior.

[87] Note also that if there is only one output good, then this will automatically hold. In this case, (10.2-9) can be rewritten as $\mathrm{TFPG}(t) = \mathrm{TP}(t) + [1 - (1/\mathrm{RS}(t))] + [Y'(t)/Y(t)]$. This expression is analogous to Equation (6.1-11) where, for the one input, one output case, we decomposed TFPG into the product of a technical progress term and a returns to scale term. In both of these equations, if output growth is positive and returns to scale are greater than one, then productivity will exceed technical progress.

and apply what they term a *semi exact estimation approach*.[88] In this approach, exact index number methods are used to greatly reduce the number of other parameters that must be estimated along with the elasticity of scale. This stream of work can be viewed as a generalization of the basic theoretical results of Diewert (1976, Lemma 2.2, equations (2.11) and Theorem 2.16), the material on noncompetitive approaches in Diewert (1978b) and additional results in Diewert (1981a, including Section 7 results on the treatment of mark-ups).

The technology of a production unit can be represented by a production, revenue or cost function. Technical progress can be conceptualized as a shift in the specified producer behavioral relationship, and returns to scale can be defined as a change in scale with the technology held fixed. Building on the work of Panzer (1989), Hall (1990) and Klette and Griliches (1996), DN4 draw attention to the fact that production, cost and revenue function based definitions of the elasticity of scale differ conceptually and are suitable for different sorts of production situations. These issues must be faced whether a traditional or a semi exact econometric approach is adopted.[89]

In the production function framework, returns to scale are defined as the percentage change in the output quantity in response to a one percent increase in each of the N input quantities. A production function framework is suitable when there is just one output, or with multiple outputs produced in fixed proportions. However, when there are multiple outputs that can be produced in varying proportions, a revenue or cost function framework may be more suitable.

When a revenue function is used to characterize the technology of the designated production unit, a measure of the elasticity of returns to scale for a multiple output, multiple input production unit can be defined conceptually as the percentage change in revenue due to a one percent increase in *each* of the input quantities. This definition of returns to scale seems problematic because most of the sources of what is referred to as returns to scale in the business and public policy literatures involve *changes* in input mix as the scale of production increases. This is the same reason why the definition of the elasticity of scale used in the data envelopment literature is problematic. According to that approach, the returns to scale measure is defined as the equiproportionate change in outputs resulting from an equiproportionate change in inputs. There is virtually no real life change in scale that does not involve changes in the input or in the output

[88] In Yoshioka, Nakajima and M. Nakamura (1994), Nakajima, M. Nakamura and Yoshioka (1998, 2001) and in a 2006 Nakajima, A. Nakamura, E. Nakamura and M. Nakamura (N4 for short) present an estimator for the elasticity of scale for a production process with multiple inputs but only one output. DN4 extend this approach to allow for multiple outputs, but with the assumption of competitive output markets and price taking behavior in these markets. Diewert and Fox (2004) generalize the approach to allow for limited types of imperfect competition and markups in output markets, building as well on Berndt and Fuss (1986), Hall (1990) and Basu and Fernald (1997). Imperfect competition in output markets is allowed for in the Hall (1990), Bartelsman (1995), and Basu–Fernald studies, but with only a single output.

[89] Other related work includes Fox (2007), Schreyer (2007), N4, Diewert and Lawrence (2005), Inklaar (2006), Balk (1998, 2001, 2003), Bartelsman (1995), Basu and Fernald (1997), Hall (1990), Morrison and Siegel (1997), and M.I. Nadiri and B. Nadiri (1999).

mix: indeed, anticipated mix changes are typically a reason for a production unit (like a nation) to strive to grow.

A cost function, like a revenue function, can be used to characterize a multi input, multi output production unit's technology. A cost function based measure of returns to scale implies a conceptually more appealing definition of returns to scale: the percentage change in cost due to a one percent increase in all output quantities. Furthermore, Diewert and Fox (2004) show that a cost function based measure of returns to scale can accommodate certain (albeit restrictive) departures from the assumption of perfectly competitive output markets. Using a cost function framework, a reciprocal form cost function based measure of the elasticity of scale is defined as the percentage change in cost due to a one percent increase in each of the output quantities, controlling for price changes.

11. Diewert–Kohli–Morrison (DKM) revenue function based productivity measures

Decompositions of a volume index number measure of overall growth into individual component sources of growth are not new; what is new are decompositions that have explicit economic interpretations. Diewert and Morrison (1986) obtain this type of economic decomposition for the Törnqvist volume index.[90] The full potential of these decompositions has only lately begun to be recognized by economists and statisticians.

In Section 5, we used the period t production function f^t to define the period t cost function, c^t. The period t production function can also be used to define the period t (net) revenue function:

$$r^t(p,x) \equiv \max_y \left\{ p \cdot y:\ y \equiv (y_1, y_2, \ldots, y_M);\ y_1 = f^t(y_2, \ldots, y_M; x) \right\}, \quad (11\text{-}1)$$

where $p \equiv (p_1, \ldots, p_M)$ is the output price vector that the producer faces and $x \equiv (x_1, \ldots, x_N)$ is the input vector.[91] Diewert and Morrison (1986) use revenue functions for period t and the comparison period s to define a family of theoretical productivity growth indexes:

$$\mathrm{RG}(p,x) \equiv r^t(p,x)/r^s(p,x). \quad (11\text{-}2)$$

[90] The same decomposition was independently derived by Kohli (1990). Diewert (2002c) obtained an analogous economic decomposition for the Fisher formula. Related material on decompositions can be found in Balk and Hoogenboom-Spijker (2003) and Diewert and Nakamura (2003).

[91] If y_m is positive (negative), then the net volume m is for an output (input). We assume that all prices p_m are positive. We assume that all input volumes x_n are positive and if the net input volume for product n is an input (output) volume, then w_n is positive (negative).

This index is the ratio of the net value of the output that can be produced using the period t versus the period s technology with input volumes held constant at some reference net input volume vector x and with prices held constant at some reference unit price vector, p. This is a different approach to the problem of controlling for total factor input utilization in judging the success of the period t versus the period s production outcomes.

Two special cases of (11-2) are of interest:

$$\begin{aligned} &RG^s \equiv RG\left(p^s, x^s\right) = r^t\left(p^s, x^s\right)/r^s\left(p^s, x^s\right) \quad \text{and} \\ &RG^t \equiv RG\left(p^t, x^t\right) = r^t\left(p^t, x^t\right)/r^s\left(p^t, x^t\right). \end{aligned} \tag{11-3}$$

The first of these, RG^s, is the theoretical productivity index obtained by letting the reference vectors p and x take on the observed period s values. The second of these, RG^t, is the theoretical productivity index obtained by letting the reference vectors be the observed period t output price vector p^t and input volume vector x^t.[92]

Under the assumption of revenue maximizing behavior in both periods, we have:

$$p^t \cdot y^t = r^t\left(p^t, x^t\right) \quad \text{and} \quad p^s \cdot y^s = r^s\left(p^s, x^s\right). \tag{11-4}$$

If these equalities hold, this means we observe values for the denominator of RG^s and the numerator of RG^t. However, we cannot directly observe the hypothetical terms, $r^t(p^s, x^s)$ and $r^s(p^t, x^t)$. The first of these is the revenue that would result from using the period t technology with the period s input volumes and output prices. The second is the revenue that would result from using the period s technology with the period t input volumes and output prices.

These hypothetical revenue figures can be inferred from observable data if we know the functional form for the period t revenue function and it is associated with an index number formula that can be evaluated with the observable data. Suppose, for example, that the revenue function has the following translog functional form:

$$\begin{aligned} \ln r^t(p, x) \equiv \alpha_s^t + \sum_{m=1}^{M} \alpha_m^t \ln p_m + \sum_{n=1}^{N} \beta_n^t \ln x_n + (1/2) \sum_{m=1}^{M} \sum_{j=1}^{M} \alpha_{mj} \ln p_m \ln p_j \\ + (1/2) \sum_{n=1}^{N} \sum_{j=1}^{N} \beta_{nj} \ln x_n \ln x_j + \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{mn} \ln p_m \ln x_n, \end{aligned} \tag{11-5}$$

where $\alpha_{mj} = \alpha_{jm}$ and $\beta_{nj} = \beta_{jn}$ and the parameters satisfy various other restrictions to ensure that $r^t(p, x)$ is linearly homogeneous in the components of the price vector p.[93]

[92] This approach can be viewed as an extension to the general N–M case of the methodology used in defining the output based measures of technical progress given in (6.1-7) and (6.1-8).

[93] These conditions can be found in Diewert (1974a, p. 139). The derivation of (6.3-1) and (6.3-2) also required the assumption of a translog technology.

Note that the coefficient vectors α_0^t, α_m^t and β_n^t can be different in each time period but that the quadratic coefficients are assumed to be constant over time.

Diewert and Morrison (1986, p. 663) show that under the above assumptions, the geometric mean of the two theoretical productivity indexes defined in (11-3) can be identified using the observable price and volume data that pertain to the two periods; i.e., we have

$$\left[\mathrm{RG}^s\,\mathrm{RG}^t\right]^{1/2} = a/(bc), \tag{11-6}$$

where a, b and c are given by

$$a \equiv p^t \cdot y^t / p^s \cdot y^s, \tag{11-7}$$

$$\ln b \equiv \sum_{m=1}^{M} (1/2)\left[\left(p_m^s y_m^s / p^s \cdot y^s\right) + \left(p_m^t y_m^t / p^t \cdot y^t\right)\right] \ln\left(p_m^t / p_m^s\right), \quad \text{and} \tag{11-8}$$

$$\ln c = \sum_{n=1}^{N} (1/2)\left[\left(w_n^s x_n^s / p^s \cdot y^s\right) + \left(w_n^t x_n^t / p^t \cdot y^t\right)\right] \ln\left(x_n^t / x_n^s\right). \tag{11-9}$$

If we have constant returns to scale production functions f^s and f^t, then the value of outputs will equal the value of inputs in each period and we have

$$p^t \cdot y^t = w^t \cdot x^t. \tag{11-10}$$

Note that the same result can be derived without the constant returns to scale assumption if we have a fixed factor that absorbs any pure profits or losses, with this fixed factor defined as in (5-18) in Section 5.

Substituting (11-10) into (11-9), we see that expression c becomes the Törnqvist input index Q_T^*. By comparing (11-8) and (3.5-2), we see also that b is the Törnqvist output price index P_T. Thus a/b is an implicit Törnqvist output volume index.

If (11-10) holds, then we have the following decomposition for the geometric mean of the product of the theoretical productivity growth indexes defined in (11-3):

$$\left[\mathrm{RG}^s\,\mathrm{RG}^t\right]^{1/2} = \left[p^t \cdot y^t / p^s \cdot y^s\right] / \left[P_T\, Q_T^*\right], \tag{11-11}$$

where P_T is the Törnqvist output price index defined in (3.5-2) and Q_T^* is the Törnqvist input volume index defined analogously to the way in which the Törnqvist output volume index is defined in (3.5-1). Diewert and Morrison (1986) use the period t and s revenue functions to define two theoretical output price effects which show how revenues would change in response to a change in a single output price:

$$P_m^s \equiv r^s\left(p_1^s, \ldots, p_{m-1}^s, p_m^t, p_{m+1}^s, \ldots, p_M^s, x^s\right) / r^s\left(p^s, x^s\right), \quad m = 1, \ldots, M, \quad \text{and} \tag{11-12}$$

$$P_m^t \equiv r^t\left(p^t, x^t\right) / r^t\left(p_1^t, \ldots, p_{m-1}^t, p_m^s, p_{m+1}^t, \ldots, p_M^t, x^t\right), \quad m = 1, \ldots, M. \tag{11-13}$$

More specifically, these theoretical indexes give the proportional changes in the value of output that would result if we changed the price of the mth output from its period s level p_m^s to its period t level p_m^t holding constant all other output prices and the input volumes at reference levels and using the same technology in both situations. For the theoretical index defined in (11-12), the reference output prices and input volumes and technology are the period s ones, whereas for the index defined in (11-13) they are the period t ones. Now define the theoretical output price effect b_m as the geometric mean of the two effects defined by (11-12) and (11-13):

$$b_m \equiv \left[P_m^s P_m^t\right]^{1/2}, \quad m = 1, \ldots, M. \tag{11-14}$$

Diewert and Morrison (1986) and Kohli (1990) show that the b_m given by (11-14) can be evaluated by the following observable expression, provided that conditions (11-4), (11-5) and (11-10) hold:

$$\ln b_m = (1/2)\left[\left(p_m^s y_m^s / p^s \cdot y^s\right) + \left(p_m^t y_m^t / p^t \cdot y^t\right)\right] \ln\left(p_m^t / p_m^s\right)$$
$$m = 1, \ldots, M. \tag{11-15}$$

Comparing (11-8) with (11-15), it can be seen that we have the following decomposition for b:

$$b = \prod_{m=1}^{M} b_m = P_T. \tag{11-16}$$

Thus the overall Törnqvist output price index, P_T, can be decomposed into a product of the individual output price effects, b_m.

Diewert and Morrison (1986) also use the period t and s revenue functions in order to define two theoretical input volume effects as follows:

$$Q_n^{*s} \equiv r^s\left(p^s, x_1^s, \ldots, x_{n-1}^s, x_n^t, x_{n+1}^s, \ldots, x_N^s\right) / r^s\left(p^s, x^s\right)$$
$$n = 1, \ldots, N, \quad \text{and} \tag{11-17}$$
$$Q_n^{*t} \equiv r^t\left(p^t, x^t\right) / r^t\left(p^t, x_1^t, \ldots, x_{n-1}^t, x_n^s, x_{n+1}^t, \ldots, x_N^t\right),$$
$$n = 1, \ldots, N. \tag{11-18}$$

These theoretical indexes give the proportional change in the value of net output that would result from changing input n from its period s level x_n^s to its period t level x_n^t, holding constant all output prices and other input volumes at reference levels and using the same technology in both situations. For the theoretical index (11-17), the reference output prices and input volumes and the technology are the period s ones, whereas for the index in (11-18) they are the period t ones.

Now define the theoretical input volume effect c_n as the geometric mean of the two effects defined by (11-17) and (11-18):

$$c_n \equiv \left[Q_n^{*s} Q_n^{*t}\right]^{1/2}, \quad n = 1, \ldots, N. \tag{11-19}$$

Diewert and Morrison (1986) show that the c_n defined by (11-19) can be evaluated by the following empirically observable expression provided that assumptions (11-4) and (11-5) hold:

$$\ln c_n = (1/2)\left[\left(w_n^s x_n^s / p^s \cdot y^s\right) + \left(w_n^t x_n^t / p^t \cdot y^t\right)\right] \ln\left(x_n^t / x_n^s\right) \tag{11-20}$$

$$= (1/2)\left[\left(w_n^s x_n^s / w^s \cdot x^s\right) + \left(w_n^t x_n^t / w^t \cdot x^t\right)\right] \ln\left(x_n^t / x_n^s\right). \tag{11-21}$$

The expression (11-21) follows from (11-20) provided that the assumptions (11-10) also hold. Comparing (11-20) with (11-9), it can be seen that we have the following decomposition for c:

$$c = \prod_{n=1}^{N} c_n \tag{11-22}$$

$$= Q_T^*, \tag{11-23}$$

where (11-23) follows from (11-22) provided that the assumptions (11-10) also hold. Thus if assumptions (11-4), (11-5) and (11-10) hold, the overall Törnqvist input volume index can be decomposed into a product of the individual input volume effects, the c_n for $n = 1, \ldots, N$.

Having derived (11-16) and (11-22), we can substitute these decompositions into (11-6) and rearrange the terms to obtain the following decomposition:

$$p^t \cdot y^t / p^s \cdot y^s = \left[\mathrm{RG}^s\, \mathrm{RG}^t\right]^{1/2} \prod_{m=1}^{M} b_m \prod_{n=1}^{N} c_n. \tag{11-24}$$

This is a decomposition of the growth in the nominal value of output into the productivity growth term $[\mathrm{RG}^s\, \mathrm{RG}^t]^{1/2}$ times the product of the output price growth effects, the b_m, times the product of the input volume growth effects, the c_n.[94] All of the effects on the right-hand side of (11-24) can be calculated using only the observable price and volume data for the two periods.[95]

12. Concluding remarks

This paper surveys the index number methods and theory behind the national productivity numbers. We close with some remarks on six aspects of the current state of

[94] An interesting case of (11-24) results when there is only one fixed input in the x vector. Then the input growth effect c_1 is unity and variable inputs appear in the y vector with negative components. The left-hand side of (11-24) becomes the pure profits ratio that is decomposed into a productivity effect times the various price effects (the b_m).

[95] See Morrison and Diewert (1990a, 1990b), Diewert (2002c), and Reinsdorf, Diewert and Ehemann (2002) for decompositions for other functional forms besides the translog. Kohli (1990), Fox and Kohli (1998), and Diewert, Lawrence and Fox (2006) use this approach to examine the factors behind the growth in the nominal GDP of several countries.

productivity measurement for nations and directions for future research. Both methodological and data challenges remain, and the two are interrelated. Better data can ease the methodology challenges.

12.1. Choice of measure effects

One goal of this paper has been to draw attention to, and help users distinguish among, different types of productivity measures for nations. We show how different ones relate to each other and to GDP per capita which is a commonly used measure of national economic well being. It is important for the differences among the measures to be kept in mind when it comes to interpreting empirical findings. Some authors make a point of helping their readers to be aware of these differences. For example, in a recent paper with important public policy implications for Canada and the US, Rao, Tang and Wang (2007) note that, unlike their earlier studies, the labor productivity measure used is for the number of persons employed rather than hours of work because they did not have comparable hours of work data by industry for the two countries. Rao, Tang and Wang note that they expect the resulting measured productivity gap with the United States to be about 10 percent higher than in their earlier studies because Canadians, on average, put in about 10 percent less hours on the job than their US counterparts. The continuing development of harmonized data for widening numbers of nations will hopefully make it possible in years to come for researchers to choose the productivity measures that best fit their applied needs rather than having to bend their analysis needs to the available data. But it will still be important for readers to be aware of how the choice of measure affects reported results.

12.2. Better price measurement = better productivity measures

The traditional index number definition of a productivity growth index is an output volume index divided by an input volume index. National statistical agencies (appropriately) collect information on output and input values and prices; not volumes and prices. Volume measures are then produced by applying price indexes to the value information about the outputs and inputs. This means that having good price statistics is critical for the quality of productivity measurement.

The new international CPI and PPI manuals provide an in-depth treatment of the theoretical, methodological, and data advances of recent decades in official consumer and producer price level measurement.[96] Nevertheless, some important problem areas remain.

The treatment of new products in price measurement remains on the critical list, and is of special relevance for productivity measurement.[97] New machinery and equipment,

[96] We are referring here to the new international Consumer Price Index Manual [T.P. Hill (2004)] and Producer Price Index Manual [Armknecht (2004)].

[97] See A. Baldwin et al. (1997), Basu et al. (2004), Greenstein (1997), R.J. Hill (1999a, 1999b, 2001, 2004), Nordhaus (1997), and Wolfson (1999). The treatment of quality change is also important. Hedonic methods are increasingly being used in this regard; see Diewert (2002d, 2003a, 2003b).

new business processes, new material inputs, and new consumer products are key ways in which technological progress is manifested. Price setting mechanisms appear to be evolving because marketing behavior is evolving to take advantage of new IT technologies, and this has potential implications for productivity measurement, along with other aspects of the treatment of new goods.[98]

The proper measurement of the prices for machinery and equipment and other capital services – that is, the proper measurement of user costs for capital – is a second important area of active debate for price measurement. In Subsection 10.1 we argued that, ideally, the measure of the user cost of capital should allow for depreciation and obsolescence effects.

For a KLEMS (capital, labor, energy, materials and services) approach, price indexes are needed as well for the intermediate inputs. The major classes of intermediate inputs at the industry level are: materials, business services and leased capital. In practice, period by period information on costs paid for a list of intermediate input categories is required along with either an intermediate input volume index or a price index for each category. There is a lack of price survey data for intermediate inputs. Price indexes for outputs are often used as proxies for the missing price indexes for intermediate inputs. Also, the intermediate input prices should, in principle, include any commodity taxes imposed on these inputs, since the tax costs are paid by producers.

Of course, many intermediate products are produced by different divisions of the same firms that use these products for producing final demand products. This intra-firm trade can be important for national productivity measurement and growth accounting when the different divisions are in different nations, in which case these movements of product will be counted as part of foreign trade.

Intermediate product transactions *among* firms can be observed and price and value statistics can be collected for these transactions as for other sorts of product transactions. This is not the case, however, for the *intra* firm transactions. As Diewert, Alterman and Eden (2007) and Mann (2007) explain, the transfer prices that firms report for intra firm transactions may not be a very satisfactory basis for the construction of price indexes for the associated flows of goods and services. Even for measuring the productivity of a single nation over time, or making inter-sector or inter-industry comparisons for productivity levels or growth within a single nation, international trade issues must be considered and dealt with. That is, international trade complicates even the choice of a measure for national level output.[99]

Finally, on the subject of relevant price measurement problem areas, there is interest not only in productivity comparisons over time, but also in inter-nation comparisons.

[98] See Hausman (2003), Hausman and Leibtag (2006, 2007), Leibtag et al. (2006), E. Nakamura and Steinsson (2006a, 2006b), Silver and Heravi (2003), Triplett (2002), Triplett and Bosworth (2004), and Timmer, Inklaar and van Ark (2005). On hedonic pricing for new goods, see Diewert, Heravi and Silver (2007).

[99] On the importance of allowing for trade in measuring productivity see, for example, Bernstein (1998), Bernstein and Mohnen (1998), Bernstein and Nadiri (1989), Diewert and Woodland (2004), Woodland and Turunen-Red (2004), Diewert (2007d), Trefler (2004), and Mann (2007).

Thus, inter-nation price statistics are needed for making inter-nation productivity comparisons: international purchasing power parity statistics (PPPs). Methodological as well as data challenges abound in the area of international price comparisons.[100]

12.3. The measurement of capital services

The OECD productivity database[101] distinguishes seven types of assets: hardware, communications equipment, other machinery, transport equipment, nonresidential buildings, structures and software. Diewert, Harrison and Schreyer (2004) state that, conceptually, there are many facets of capital input that bear a direct analogy to labor input. Capital goods are seen as carriers of capital services that constitute the actual input in the production process just as workers are seen as carriers of labor services. When rentals and the cost of capital services cannot be observed directly, methods must be adopted to approximate the costs of capital services. Much progress has been made on the measurement of capital services and user costs, but much still remains to be done.

For example, Brynjolfsson and Hitt (2000)[102] and Corrado, Hulten and Sichel (2006) note that both firm level and national income accounting practice has historically treated expenditure on intangible inputs such as software and R&D as an intermediate expense and not as an investment that is part of GDP. Corrado, Hulten and Sichel (2006) find that the inclusion of intangibles makes a significant difference in the measured pattern of economic growth: the growth rates of output and of output per worker are found to increase at a faster rate when intangibles are included than under the status quo case in which intangible capital is ignored.

Brynjolfsson and Hitt (2000) argue there are important interaction effects between the intangible business practice and process capital investments that are going unmeasured. Brynjolfsson and Hitt (1995), Atrostic et al. (2004), Atrostic and Nguyen (2007), and Dufour, Nakamura and Tang (2007) provide empirical evidence for multiple countries suggesting that IT-users that also invested in organizational capital had higher gains in productivity compared to firms that invested only in IT capital or only in adopting

[100] See, for instance, Armstrong (2001, 2003, 2007), Balk (1996), Diewert (2000, 2005d, 2006b), R.J. Hill (1999a, 1999b, 2001, 2004, 2007), R.J. Hill and Timmer (2004), and D.S.P Rao (2007).

[101] See OECD (2005).

[102] In a 2000 *Journal of Economic Perspectives* article, Brynjolfsson and Hitt write: "Changes in multifactor productivity growth, in turn, depend on accurate measures of final output. However, nominal output is affected by whether firm expenditures are expensed, and therefore deducted from value added, or capitalized and treated as investment. As emphasized throughout this paper, information technology is only a small fraction of a much larger complementary system of tangible and intangible assets. However, current statistics typically treat the accumulation of intangible capital assets, such as new business processes, new production systems and new skills, as expenses rather than as investments. This leads to a lower level of measured output in periods of net capital accumulation. Second, current output statistics disproportionately miss many of the gains that information technology has brought to consumers such as variety, speed, and convenience . . .". See also Berndt and Morrison (1995), Brynjolfsson and Hitt (1995), Colecchia and Schreyer (2001, 2002), and Prud'homme, Sanga and Yu (2005).

new business practices. The idea is that investments in tangible and intangible assets reinforce one another.

Findings of complementarities between business practices and high tech processes lend support to the proposition that there are also important complementarities between the largely intangible unmeasured assets of firms and the (mostly tangible) assets that are being measured by national statistics agencies.

Regarding the measured assets, Jorgenson argues that rental values should be imputed on the basis of estimates of capital stocks and of property compensation rates, with the capital stock at each point of time represented as a weighted sum of past investment. The weights are viewed as measures of the relative efficiencies of capital goods of different ages and of the compensation received by the owners.[103] While agreeing with the objective of adopting a user cost approach for asset pricing, nevertheless it is important to note that the theoretical and empirical basis is slim for many of the practical choices that must be made in doing this. Substantial differences in the productivity measurement results can result from different choices about things such as physical depreciation rates for which empirical or other scientific evidence is largely lacking. Nomura and Futakamiz (2005) report on an initiative in Japan to use the complete records of assets in place in some companies and also the registration data for particular assets to determine the service life of individual assets. We strongly support this sort of data collection initiative.

Yet another capital measurement issue is that the System of National Accounts (SNA) does not regard the placing of nonproduced assets at the disposal of a producer as production in itself but as an action giving rise to property income. Nomura (2004, Chapter 4) shows that neglecting land and inventories leads to a reduction in the average TFP growth rate.

12.4. Labor services of workers and service products

Diewert, Harrison and Schreyer (2004) state that, conceptually, there are many facets of capital input that bear a direct analogy to labor input. They also state that when rentals and the cost of capital services cannot be observed directly, methods must be adopted to approximate the costs of capital services. With hourly workers, we have data on their "rental rates". But is it the *right* price information?

The issue of what sorts of labor to count must be agreed on first. For productivity measurement purposes, there is agreement that the labor to be counted is what is used for the production within the boundary of the System of National Accounts (SNA). However, this still leaves three alternative definitions of the hours of work that have been the subject of international debate on the proper measurement of labor input:

(H1) *Active production time*.

[103] See for example Jorgenson (1963, 1980, 1995a), and Christensen and Jorgenson (1969, 1970).

(H2) *Paid hours*, including agreed on allocations of employer financed personal time, some of which may be taken away from the work site like paid vacation or sick days.
(H3) *Hours at work* whether or not they are paid hours or used for production.

Each of these measurement concepts implies a different factorization of the dollar amount spent on labor into quantity and price components.

The productivity programs in Canada and the United States use the concept of hours at work: the H3 concept. H3 is also the concept that is agreed on for the 1993 System of National Accounts as the most appropriate measure for determining the volume of work.

For many workers, time at work includes some hours in addition to what are stipulated in formal employment agreements. These are volunteered or informally coerced unpaid overtime hours. Hours at work (H3) could rise (or fall) with or without changes in paid hours (H2), but it seems unlikely there would be changes in hours at work without corresponding changes in the same direction in active production time (H1). Conceptually at least, the H3 concept includes the increasing amounts of work done at home by those tele-commuting, from other nations.[104]

A different sort of labor input measurement issue is that none of the measures discussed take account of differences in the knowledge and skills and innate abilities vested in workers.

Economic development history could be written as the progressive substitution of machine for human services. Farmer laborers tilling the fields were replaced by machines that do the tilling, pulled by farmers riding tractors. Phone operators were replaced by electronic routing systems. Type setters were replaced by automated printing processes. Secretarial typing of research papers was replaced by word processors on home or portable computers and the typing of the researchers themselves. To properly account for these substitutions of these sorts, we would need data on the respective volumes and prices, or the value figures and prices, for workers with different types and levels of specific skills.

12.5. A need for official statistics and business world harmonization

The models that economists use to interpret TFPG estimates typically rule out many of the ways in which business and government leaders attempt to raise productivity. The dominant economic index number approach is built, to date, on a neoclassical foundation assuming perfect competition, perfect information and, in most studies, constant returns to scale.

[104] Note too that for materials, inventories represent the difference between the paid quantity for a given time period (concept H2) and the quantity used in that time period (concept H1). In contrast, for material inputs, time "at work" (quantity concept H3) is the same as the counterpart of "paid time" because the materials are owned continuously, once paid for. Diewert and Nakamura (1999) are silent on the issue of inventories for diesel and lube oil.

In a world where all factor inputs are paid their marginal products and there is no potential for reaping increasing returns to scale, then the only way in which growth in output could occur would be through increased input use or through changes in external circumstances, including spillovers from the R&D of others.[105] This is the world assumed by Solow (1957), for example. For such a world, after removing all factor costs in evaluating productivity growth, we would be left with only revenue growth due to *purely external factors*. Thus, Jorgenson (1995a) writes:

> "The defining characteristic of productivity as a source of economic growth is that the incomes generated by higher productivity are external to the economic activities that generate growth" (p. xvii).

However, this definition of productivity growth seems unlikely to satisfy Harberger's (1998, p. 1) recommendation that we should approach the measurement of productivity by trying to "think like an entrepreneur or a CEO, or a production manager". What CEO would announce a productivity improvement plan for their company, and then add that it depends entirely on external happenings including spillovers from the R&D of competitors?

The private business sector is the engine of productivity growth in capitalist economies. Business and government leaders need to be able to communicate effectively about economic policy issues. National productivity matters are of concern to both business and government leaders. When asked, business leaders mostly report that they make little or no use of the productivity measures of economists, preferring instead to use single factor input–output type performance measures. However, businesses make ubiquitous use of real revenue/cost ratios, and we have shown that this is one way of writing a TFPG index. We suggest that the main differences between the way that economists and business leaders have traditionally thought about productivity lie in defining technical change as disembodied, the assumption of constant returns to scale, and the definition of productivity change as due to externalities. Business leaders see technological change as largely embodied in machines, business practices and people working for them. They are obsessed with finding ways to profit from various sorts of returns to scale. They see increasing productivity as a core function of business managers and productivity gains as being achieved by economic activities that also generate growth. And they definitely intend to capture as much as possible of the incomes generated by the productivity gains of their companies.

At present, there is a serious conceptual gulf between the economic approach to the interpretation of TFPG measures and the business world perception of what productivity growth is. This is unfortunate since it is the private business sector on which nations must mostly rely for their economic growth. The challenge for index number theorists

[105] Studies of TFPG focusing explicitly on externalities such as R&D spillovers include Bernstein and Nadiri (1989), Bernstein (1996), and Jaffe (1986). Bernstein (1998) and Bernstein and Mohnen (1998) extend the theory and empirical treatment of spillover effects on productivity growth to an international context.

is to develop models that incorporate rather than assume away what economic practitioners view as some of the main means by which total factor productivity improvement is accomplished. One key to making headway on this goal may be to notice that the productivity measures themselves can be computed without making any of the restrictive assumptions that have been used in showing that certain of these measures can also be derived from economic optimizing models.

12.6. The role of official statistics for globally united nations

Masahiro Kuroda (2006), in his capacity as Director of the Economic and Social Research Institute of the Cabinet Office of the Government of Japan, calls attention to the need in Japan and elsewhere, to review and change the legal framework for official statistics. He links this need to, for example, an emerging need for new rules to enable wider use of administrative records as well as the survey data collected by governments.[106] We heartily support the sorts of goals enunciated by Kuroda. There is an urgent need for initiatives that will allow statistical agencies to continue to produce more and better data, and this situation will inevitably bring into play cost pressures. In addition to the long established importance of official statistics in the management of national economies, official statistics have been evolving into an important medium for communication both within and among nations. Increasingly, the national choices that affect all of us as global inhabitants are being made in the context, and with the aid, of official statistics, including measures of the productivity of nations.

References

Abramovitz, M. (1956). "Resource and output trends in the United States since 1870". American Economic Review 46 (2), 5–23.

Ahmad, N., Lequiller, F., Marianna, P., Pilat, D., Schreyer, P., Wolfl, A. (2003). "Comparing labour productivity growth in the OECD area: The role of measurement". OECD STI Working Paper 2003/14. OECD, Paris.

Allen, R.C., Diewert, W.E. (1981). "Direct versus implicit superlative index number formulae". Review of Economics and Statistics 63 (3), 430–435; reprinted as Chapter 10 in Diewert and Nakamura (1993, pp. 275–286).

Allen, R.G.D. (1949). "The economic theory of index numbers". Economica N. S. 16, 197–203.

Armitage, H.M., Atkinson, A.A. (1990). The Choice of Productivity Measures in Organizations. The Society of Management Accountants, Hamilton, Ontario.

Armknecht, P.A. (Ed.) (2004). Producer Price Index Manual: Theory and Practice (PPI Manual). International Labour Office, International Monetary Fund, Organisation for Economic Co-operation and Development, Eurostat, United Nations, and The World Bank. Chapters and whole can be downloaded for free at http://www.imf.org/external/np/sta/tegppi/index.htm.

[106] Halliwell (2005) provides thought provoking perspectives and possible ways of proceeding. For other related issues see Diewert (2006a), Nakamura and Diewert (1996) and McMahon (1995).

Armstrong, K.G. (2001). "What impact does the choice of formula have on international comparisons?". Canadian Journal of Economics 34 (3), 697–718 (August).
Armstrong, K.G. (2003). "A restricted-domain multilateral test approach to the theory of international comparisons". International Economic Review 44 (1), 31–86 (February).
Armstrong, K.G. (2007). "Hyperextended-real-valued indexes of absolute dissimilarity". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 4: International Comparisons and Trade. Trafford Press. Chapter 11.
Aschauer, D.A. (1989). "Public investment and productivity growth in the group of seven". Economic Perspectives 13 (1), 17–25.
Atkinson, A., Kaplan, R.S., Young, S.M. (1995). Management Accounting. Prentice Hall, Englewood Cliffs, NJ.
Atrostic, B.K., Bough-Nielsen, P., Motohashi, K., Nguyen, S.V. (2004). "IT, productivity and growth in enterprise: New results from international data". In: The Economic Impact of ICT: Measurement, Evidence and Implications. OECD, Paris.
Atrostic, B.K., Nguyen, S.V. (2006). "Computer investment, computer networks, and productivity". In: Hulten, C., Berndt, E. (Eds.), Hard-to-Measure Goods and Services: Essays in Memory of Zvi Griliches. University of Chicago Press. In press.
Atrostic, B.K., Nguyen, S.V. (2007). "IT and business process impacts on US plant-level productivity". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 8: ICT and Business Process Effects. Trafford Press. Chapter 3. In press.
Baily, M.N. (1981). "The productivity growth slowdown and capital accumulation". American Economic Review 71, 326–331.
Baldwin, A., Duspres, P., Nakamura, A., Nakamura, M. (1997). "New goods from the perspective of price index making in Canada and Japan". In Bresnahan and Gordon (1997, pp. 437–474).
Baldwin, J.R., Harchauoui, T.M., Hosein, J., Maynard, J.-P. (2001). "Productivity: Concepts and trends". In: Baldwin, J.R., Beckstead, D., Dhaliwal, N., Durand, R., Gaudreault, V., Harchaoui, T.M., Hosein, J., Kaci, M., Maynard, J.-P. (Eds.), Productivity Growth in Canada, Catalogue no. 15-204-XPE. Statistics Canada, Ottawa, pp. 51–60.
Baldwin, J.R., Jarmin, R., Tang, J. (2004). "Small North American producers give ground in the 1990s". Small Business Economics 23 (4), 349–361.
Baldwin, J.R., Maynard, J.-P., Tanguay, M., Wong, F., Yan, B. (2005). "A comparison of Canadian and US productivity levels: An exploration of measurement issues". Paper 11F0027MIE No. 028. Statistics Canada, Ottawa.
Baldwin, J.R., Maynard, J.-P., Wong, F. (2005). "The output gap between Canada and the United States: The role of productivity (1994–2002)". In: Analytical Paper Catalogue no. 11-624-MIE – No. 009. Statistics Canada, Ottawa.
Baldwin, J.R., Tanguay, M. (2006). "Estimating depreciation rates for the productivity accounts". Micro-Economic Analysis Division, Statistics Canada, Ottawa.
Balk, B.M. (1995). "Axiomatic price theory: A survey". International Statistical Review 63 (1), 69–93.
Balk, B.M. (1996). "A comparison of ten methods for multilateral international price and volume comparisons". Journal of Official Statistics 12, 199–222.
Balk, B.M. (1998). Industrial Price, Quantity and Productivity Indices. Kluwer Academic Publishers, Boston, MA.
Balk, B.M. (2000). "Divisia price and quantity indices: 75 years after". Department of Statistical Methods, Statistics Netherlands, PO Box 4000, 2270 JM Voorburg, The Netherlands.
Balk, B.M. (2001). "Scale efficiency and productivity change". Journal of Productivity Analysis 15, 159–183.
Balk, B.M. (2003). "The residual: On monitoring and benchmarking firms, industries, and economies with respect to productivity". Journal of Productivity Analysis 20, 5–47.
Balk, B.M., Hoogenboom-Spijker, E. (2003). "The measurement and decomposition of productivity change: Exercises on the Netherlands' manufacturing industry". Discussion paper 03001. Statistics Netherland. Available for download at http://www.nationalstatistics.gov.uk/events/CAED/abstracts/downloads/hoogenboom-spijker.pdf.

Bartelsman, E.J. (1995). "Of empty boxes: Returns to scale revisited". Economics Letters 49, 59–67.
Bartelsman, E.J., Doms, M. (2000). "Understanding productivity: Lessons from longitudinal microdata". Journal of Economic Literature, 569–594 (September).
Basu, S., Fernald, J.G. (1997). "Returns to scale in US production: Estimates and implications". Journal of Political Economy 105 (2), 249–283.
Basu, S., Fernald, J.G., Oulton, N., Srinivasan, S. (2004). "The case of the missing productivity growth, or does information technology explain why productivity accelerated in the United States but not in the United Kingdom?". In: Gertler, M., Rogoff, K. (Eds.), 2003 NBER Macroeconomics Annual, vol. 6. MIT Press, Cambridge, MA, pp. 9–63.
Berndt, E.R. (1991). The Practice of Econometrics Classic and Contemporary. Addison–Wesley, New York.
Berndt, E.R., Fuss, M.A. (1986). "Productivity measurement with adjustments for variations in capacity utilization and other forms of temporary equilibrium". Journal of Econometrics 33, 7–29.
Berndt, E.R., Khaled, M.S. (1979). "Parametric productivity measurement and choice among flexible functional forms". Journal of Political Economy 87 (61), 1220–1245.
Berndt, E.R., Morrison, C.J. (1995). "High-tech capital formation and economic performance in US manufacturing industries: An exploratory analysis". Journal of Econometrics 65, 9–43.
Berndt, E.R., Wood, D.O. (1975). "Technology, prices and the derived demand for energy". Review of Economics and Statistics, 259–268 (August).
Bernstein, J.I. (1996). "The Canadian communication equipment industry as a source of R&D spillovers and productivity growth". In: Howitt, P. (Ed.), The Implications of Knowledge-Based Growth for Micro-Economic Policies. University of Calgary Press, Calgary, Alberta.
Bernstein, J.I. (1998). "Factor intensities, rates of return and international R&D spillovers: The case of Canadian and US industries". Annales d'Economie et de Statistique 49/50, 541–564.
Bernstein, J.I. (1999). "Total factor productivity growth in the Canadian life insurance industry". Canadian Journal of Economics 32 (20), 500–517.
Bernstein, J.I., Mohnen, P. (1998). "International R&D spillovers between US and Japanese R&D intensive sectors". Journal of International Economics 44, 315–338.
Bernstein, J.I., Nadiri, M.I. (1989). "Research and development and intraindustry spillovers: An application of dynamic duality". Review of Economic Studies 1989, 249–267 (April).
Black, S.E., Lynch, L.M. (1996). "Human-capital investments and productivity". American Economic Review 86 (2), 263–267.
Böhm-Bawerk, E. von (1891). The Positive Theory of Capital. G.E. Stechert, New York. W. Smart translator of the original German book published in 1888.
Boskin, M.J. (1997). "Some thoughts on improving economic statistics to make them more relevant in the information age". Document prepared for the Joint Economic Committee, Office of the Vice Chairman, United States Congress. Government Printing Office, Washington DC.
Bresnahan, T., Brynjolfsson, E., Hitt, L. (2002). "Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence". Quarterly Journal of Economics 117 (1), 339–376.
Bresnahan, T.F., Gordon, R.J. (1997). The Economics of New Goods. The University of Chicago Press, Chicago, IL.
Brynjolfsson, E., Hitt, L. (1995). "Information technology as a factor of production: The role of differences among firms". Economics of Innovation and New Technology.
Brynjolfsson, E., Hitt, L. (2000). "Beyond computation: Information technology, organizational transformation and business performance". Journal of Economic Perspectives 14 (4), 23–48 (Fall).
Burgess, D.F. (1974). "A cost minimization approach to import demand equations". Review of Economics and Statistics 56 (2), 224–234.
Caves, D.W., Christensen, L., Diewert, W.E. (1982a). "Multilateral comparisons of output, input, and productivity using superlative index numbers". Economic Journal 92 (365), 73–86.
Caves, D.W., Christensen, L., Diewert, W.E. (1982b). "The economic theory of index numbers and the measurement of input, output, and productivity". Econometrica 50 (6), 1393–1414.
Christensen, L.R., Jorgenson, D.W. (1969). "The measurement of US real capital input, 1929–1967". Review of Income and Wealth 15 (4), 293–320.

Christensen, L.R., Jorgenson, D.W. (1970). "US real product and real factor input, 1929–1967". Review of Income and Wealth 16 (1), 19–50.
Christensen, L.R., Jorgenson, D.W. (1973). "Measuring the performance of the private sector of the US economy, 1929–1969". In: Moss, M. (Ed.), Measuring Economic and Social Performance. Columbia University Press, New York, pp. 233–351.
Christensen, L.R., Jorgenson, D.W., Lau, L.J. (1971). "Conjugate duality and the transcendental logarithmic production function". Econometrica 39, 255–256.
Christensen, L.R., Jorgenson, D.W., Lau, L.J. (1973). "Transcendental logarithmic production frontiers". Review of Economics and Statistics 55 (1), 28–45.
Church, A.H. (1909). "Organisation by production factors". The Engineering Magazine 38, 184–194.
Clemhout, S. (1963). "The ratio method of productivity measurement". Economic Journal 73 (290), 358–360. June.
Colecchia, A., Schreyer, P. (2001). "The impact of information communications technology on output growth". STI Working Paper 2001/7. OECD, Paris.
Colecchia, A., Schreyer, P. (2002). "The contribution of information and communications technologies to economic growth in nine countries". In: OECD Economic Studies, vol. 34. OECD, Paris, pp. 153–171.
Cooper, R., Haltiwanger, J. (1993). "The aggregate implications of machine replacement: Theory and evidence". American Economic Review 83, 360–382.
Copeland, M.A. (1937). "Concepts of national income". In: Studies in Income and Wealth, vol. 1. NBER, New York, pp. 3–63.
Corrado, C., Hulten, C., Sichel, D. (2006). "Intangible capital and economic growth". Finance and Economics Discussion Series 2006-24. US Board of Governors of the Federal Reserve System. http://www.federalreserve.gov/pubs/feds/2006/200624/200624abs.html.
de Haan, M., Balk, B.M., Bergen van den, D., de Heij, R., Langenberg, H., Zijlmans, G. (2005). "The development of productivity statistics at statistics Netherlands". OECD Workshop on Productivity Measurement, Madrid, 17–19 October.
Dean, E.R., Harper, M.J. (2000). The BLS Productivity Measurement Program. US Bureau of Labor Statistics, Washington.
Denison, E.F. (1967). Why Growth Rates Differ: Post-War Experience in Nine Western Countries. Brookings Institution, Washington, DC.
Denison, E.F. (1979). Accounting for Slower Economic Growth. Brookings Institution, Washington, DC.
Denny, M., Fuss, M., Waverman, L. (1981). "The measurement and interpretation of total factor productivity in regulated industries, with an application to Canadian telecommunications". In: Cowing, T., Stevenson (Eds.), Productivity Measurement in Regulated Industries. Academic Press, New York, pp. 179–218.
Diewert, W.E. (1969). "Functional form in the theory of production and consumer demand". PhD Dissertation. University of California at Berkeley.
Diewert, W.E. (1971). "An application of the Shephard duality theorem, a generalized Leontief production function". Journal of Political Economy 79 (3), 481–507.
Diewert, W.E. (1973). "Functional forms for profit and transformation functions". Journal of Economic Theory 6 (3), 284–316.
Diewert, W.E. (1974a). "Applications of duality theory". In: Intrilligator, M.D., Kendrick, D.A. (Eds.), Frontiers of Quantitative Economics, vol. II. North-Holland Publishing Co., Amsterdam, pp. 106–171.
Diewert, W.E. (1974b). "Functional forms for revenue and factor requirement functions". International Economic Review 15, 119–130.
Diewert, W.E. (1976). "Exact and superlative index numbers". Journal of Econometrics 4 (2), 115–146; reprinted as Chapter 8 in Diewert and Nakamura (1993, pp. 223–252).
Diewert, W.E. (1977). "Walras theory of capital formation and the existence of a temporary equilibrium". In: Schwödiauer, G. (Ed.), Equilibrium and Disequilibrium in Economic Theory. D. Reidel, Dordrecht, pp. 73–126.
Diewert, W.E. (1978a). "Hicks' aggregation theorem and the existence of a real value added function". In: Fuss, M., McFadden, D. (Eds.), Production Economics: A Dual Approach to Theory and Applications,

vol. 2. North-Holland, Amsterdam, pp. 17–52; reprinted as Chapter 15 in Diewert and Nakamura (1993, pp. 435–470).

Diewert, W.E. (1978b). "Superlative index numbers and consistency in aggregation". Econometrica 46, 883–900; reprinted as Chapter 9 in Diewert and Nakamura (1993, pp. 253–273).

Diewert, W.E. (1980a). "Aggregation problems in the measurement of capital". In: Usher, D. (Ed.), The Measurement of Capital. University of Chicago Press, Chicago, pp. 433–528.

Diewert, W.E. (1980b). "Hicks' aggregation theorem and the existence of a real value added function". In: Fuss, M., McFadden, D. (Eds.), Production Economics: A Dual Approach to Theory and Applications, vol. 2. North-Holland, Amsterdam, pp. 17–51.

Diewert, W.E. (1981a). "The economic theory of index numbers: A survey". In: Deaton, A. (Ed.), Essays in the Theory and Measurement of Consumer Behaviour in Honour of Sir Richard Stone. Cambridge University Press, London, pp. 163–208; reprinted as Chapter 7 in Diewert and Nakamura (1993, pp. 177–221).

Diewert, W.E. (1981b). "The theory of total factor productivity measurement in regulated industries". In: Cowing, T.G., Stevenson, R.E. (Eds.), Productivity Measurement in Regulated Industries. Academic Press, New York, pp. 17–44.

Diewert, W.E. (1982). "Duality approaches to microeconomic theory". In: Arrow, K.J., Intriligator, M.D. (Eds.), Handbook of Mathematical Economics, vol. II. North-Holland, Amsterdam, pp. 535–599, is an abridged version of a 1978 Stanford University technical report by Diewert; a more complete abridged version is published as Chapter 6 in Diewert and Nakamura (1993, pp. 105–175).

Diewert, W.E. (1983). "The theory of the output price index and the measurement of real output change". In: Diewert, W.E., Montmarquette, C. (Eds.), Price Level Measurement. Statistics Canada, Ottawa, pp. 1049–1113. Reprinted in: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 6: Index Number Theory. Trafford Press. Chapter 12. In press.

Diewert, W.E. (1987). "Index numbers". In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: A Dictionary of Economics, vol. 2. Macmillan, London, pp. 767–780; reprinted as Chapter 5 in Diewert and Nakamura (1993, pp. 71–104).

Diewert, W.E. (1988). "Test approaches to international comparisons". Measurement in Economics, 67–86; reprinted as Chapter 12 in Diewert and Nakamura (1993, pp. 67–86).

Diewert, W.E. (1992a). "The measurement of productivity". Bulletin of Economic Research 44 (3), 163–198.

Diewert, W.E. (1992b). "Fisher ideal output, input, and productivity indexes revisited". Journal of Productivity Analysis 3, 211–248; reprinted as Chapter 13 in Diewert and Nakamura (1993, pp. 211–248).

Diewert, W.E. (1993a). "Overview of volume I". In Diewert and Nakamura (1993, pp. 1–31).

Diewert, W.E. (1993b). "The early history of price index research". In Diewert and Nakamura (1993, pp. 33–66).

Diewert, W.E. (1993c). "Symmetric means and choice under uncertainty". In Diewert and Nakamura (1993, pp. 355–434).

Diewert, W.E. (1995). "Functional form problems in modeling insurance and gambling". The Geneva Papers on Risk and Insurance Theory 20, 135–150.

Diewert, W.E. (1998a). "Index number issues in the consumer price index". The Journal of Economic Perspectives 12 (1), 47–58.

Diewert, W.E. (1998b). "High inflation, seasonal commodities and annual index numbers". Macroeconomic Dynamics 2, 456–471.

Diewert, W.E. (1999). "Axiomatic and economic approaches to international comparisons". In: Heston, A., Lipsey, R.E. (Eds.), International and Interarea Comparisons of Income, Output and Prices. In: Studies in Income and Wealth, vol. 61. University of Chicago Press, Chicago, pp. 13–87.

Diewert, W.E. (2000). "Alternative approaches to measuring productivity and efficiency". Paper prepared for the North American Productivity Workshop at Union College, Schenectady, New York, June 15–17.

Diewert, W.E. (2001a). "Which (old) ideas on productivity measurement are ready to use?". In: Hulten, C.R., Dean, E.R., Harper, M.J. (Eds.), New Developments in Productivity Analysis. In: National Bureau of Economic Analysis Studies in Income and Wealth, vol. 63. The University of Chicago Press, Chicago, pp. 85–101.

Diewert, W.E. (2001b). "Measuring the price and quantity of capital services under alternative assumptions". Discussion Paper. Department of Economics, University of British Columbia.
Diewert, W.E. (2001c). "Productivity growth and the role of government". Discussion paper. Department of Economics, University of British Columbia. http://www.econ.ubc.ca/discpapers/dp0113.pdf.
Diewert, W.E. (2002a). "Productivity trends and determinants in Canada". In: Rao, S., Sharpe, A. (Eds.), Productivity Issues in Canada. University of Calgary Press, pp. 31–58.
Diewert, W.E. (2002b). "Harmonized indexes of consumer prices: Their conceptual foundations". Swiss Journal of Economics and Statistics 138 (4), 547–637.
Diewert, W.E. (2002c). "The quadratic approximation lemma and decompositions of superlative indexes". Journal of Economic and Social Measurement 28, 63–88.
Diewert, W.E. (2002d). "Hedonic regressions: A consumer theory approach". In: Feenstra, R.C., Shapiro, M.D. (Eds.), Scanner Data and Price Indexes. In: Studies in Income and Wealth, vol. 64. NBER and University of Chicago Press, pp. 317–348. http://www.econ.ubc.ca/diewert/scan.pdf.
Diewert, W.E. (2003a). "The treatment of owner occupied housing and other durables in a consumer price index". Discussion Paper 03-08. Department of Economics, University of British Columbia, Vancouver, Canada. http://www.econ.ubc.ca/discpapers/dp0308.pdf.
Diewert, W.E. (2003b). "Hedonic regressions: A review of some unresolved issues". Paper presented at the 7th Meeting of the Ottawa Group, Paris, May 27–29. http://www.ottawagroup.org/pdf/07/Hedonics%20unresolved%20issues%20-%20Diewert%20(2003).pdf.
Diewert, W.E. (2004a). "Measuring capital". Discussion Paper 04–10. Department of Economics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z1.
Diewert, W.E. (2004b). "Durables and user costs". Chapter 23 in Hill (2004, pp. 419–441).
Diewert, W.E. (2004c). "On the stochastic approach to linking the regions in the ICP". Discussion Paper 04-16. Department of Economics, University of British Columbia, November. http://www.econ.ubc.ca/diewert/icp.pdf.
Diewert, W.E. (2005a). "On measuring inventory change in current and constant dollars". Discussion Paper 05-12. Department of Economics, University of British Columbia, Vancouver, Canada, August. http://www.econ.ubc.ca/discpapers/dp0512.pdf.
Diewert, W.E. (2005b). "The measurement of capital: Traditional user cost approaches". Chapter 1 in The Measurement of Business Capital, Income and Performance, Tutorial presented at the University Autonoma of Barcelona, Spain, September 21–22, 2005; revised December, 2005. http://www.econ.ubc.ca/diewert/barc1.pdf.
Diewert, W.E. (2005c). "Issues in the measurement of capital services, depreciation, asset price changes and interest rates". In: Corrado, C., Haltiwanger, J., Sichel, D. (Eds.), Measuring Capital in the New Economy. University of Chicago Press, Chicago, pp. 479–542.
Diewert, W.E. (2005d). "Weighted country product dummy variable regressions and index number formulae". Review of Income and Wealth Series 51 (4), 561–571 (December).
Diewert, W.E. (2005e). "Some issues concerning index number theory". Paper presented at the ESRI "Conference on the Next Steps for the Japanese System of National Accounts: Towards More Accurate Measurement and More Comprehensive Accounts", held in Tokyo, March 24-25, 2005. http://www.esri.go.jp/jp/workshop/050325/050325paper09.pdf.
Diewert, W.E. (2005f). "Welfare, productivity and changes in the terms of trade". Paper presented at the US Bureau of Economic Analysis, November 22.
Diewert, W.E. (2006a). "Services and the new economy: Data needs and challenges". In: Lipsey, R.E., Nakamura, A.O. (Eds.), Services Industries and the Knowledge Based Economy. University of Calgary Press, Calgary, Alberta, pp. 557–581. Chapter 15.
Diewert, W.E. (2006b). "Similarity indexes and criteria for spatial linking". In: Rao, D.S.P. (Ed.), Purchasing Power Parities of Currencies: Recent Advances in Methods and Applications. Edward Elgar, Cheltenham, UK, pp. 155–176. Chapter 8.
Diewert, W.E. (2006c). "Comment on 'Aggregation issues in integrating and accelerating BEA's accounts: Improved methods for calculating GDP by industry' ". In: Jorgenson, D.W., Landefeld, J.S., Nordhaus,

W.D. (Eds.), A New Architecture for the US National Accounts. University of Chicago Press for the Conference on Research in Income and Wealth (CRIW).
Diewert, W.E. (2006d). "The measurement of income". In: The Measurement of Business Capital, Income and Performance (Chapter 7), Tutorial presented at the University Autonoma of Barcelona, Spain, September 21–22, 2005; revised February 2006. http://www.econ.ubc.ca/diewert/barc7.pdf.
Diewert, W.E. (2007a). "The measurement of capital". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 9: Capital and Income. Trafford Press. Chapter 2.
Diewert, W.E. (2007b). "Index numbers". In: Durlauf, S., Blume, L. (Eds.), New Palgrave Dictionary of Economics, Palgrave Macmillan. In press. Working paper version at http://www.econ.ubc.ca/diewert/dp0702.pdf.
Diewert, W.E. (2007c). "On the stochastic approach to index numbers". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 4: International Comparisons and Trade. Trafford Press. Chapter 14.
Diewert, W.E. (2007d). "Changes in the terms of trade and Canada's productivity performance". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 4: International Comparisons and Trade. Trafford Press. Chapter 2.
Diewert, W.E., Alterman, W.F., Eden, L. (2007). "Transfer prices and import and export price indexes: Theory and practice". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 4: International Comparisons and Trade. Trafford Press. Chapter 4.
Diewert, W.E., Fox, K.J. (1999). "Can measurement error explain the productivity paradox?". Canadian Journal of Economics 32 (2), 251–281.
Diewert, W.E., Fox, K.J. (2004). "On the estimation of returns to scale, technical progress and monopolistic markups". Working paper. Department of Economics, University of British Columbia. http://www.econ.ubc.ca/discpapers/dp0409.pdf.
Diewert, W.E., Harrison, A., Schreyer, P. (2004). "Cost of capital services in the production account". Paper presented to the meeting of the Canberra Group in London, September.
Diewert, W.E., Heravi, S., Silver, M. (2007). "Hedonic imputation indexes versus time dummy hedonic indexes". In: Diewert, W.E., Greenless, J., Hulten, C. (Eds.), Price Index Concepts and Measurement. In: NBER Studies in Income and Wealth. University of Chicago Press. In press.
Diewert, W.E., Lawrence, D.A. (2000). "Progress in measuring the price and quantity of capital". In: Lau, L.J. (Ed.), Econometrics, Volume 2: Econometrics and the Cost of Capital: Essays in Honor of Dale W. Jorgenson. MIT Press, Cambridge, MA, pp. 273–326.
Diewert, W.E., Lawrence, D.A. (2005). "Australia's productivity growth and the role of information and communications technology: 1960–2004". Report prepared by Meyrick and Associates for the Department of Communications, Information Technology and the Arts, Canberra.
Diewert, W.E., Lawrence, D.A., Fox, K.J. (2006). "The contributions of productivity, price changes and firm size to profitability". Journal of Productivity Analysis 26 (1), 1–13.
Diewert, W.E., Mizobuchi, H., Nomura, K. (2007). "The contribution of the market sector to changes in Japan's living standards". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 7: Productivity Performance. Trafford Press. Chapter 11.
Diewert, W.E., Morrison, C.J. (1986). "Adjusting output and productivity indexes for changes in the terms of trade". Economic Journal 96, 659–679.
Diewert, W.E., Nakajima, T., Nakamura, A., Nakamura, E., Nakamura, M. (also referred to as DN4) (2007). "The definition and estimation of returns to scale with an application to Japanese industries". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 7: Productivity Performance. Trafford Press. Chapter 11.
Diewert, W.E., Nakamura, A.O. (1993). Essays in Index Number Theory, vol. I. North-Holland, Amsterdam.
Diewert, W.E., Nakamura, A.O. (1999). "Benchmarking and the measurement of best practice efficiency: An electricity generation application". Canadian Journal of Economics 32 (2), 570–588.
Diewert, W.E., Nakamura, A.O. (2003). "Index number concepts, measures and decompositions of productivity growth". Journal of Productivity Analysis 19 (2/3), 127–160.

Diewert, W.E., Nakamura, A.O., Schreyer, P. (2007). "Capitalized net product as a discrete time proxy for discounted optimal consumption". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 9: Capital and Income. Trafford Press. Chapter 8.
Diewert, W.E., Schreyer, P. (2006a). "The measurement of capital". In: Durlauf, S., Blume, L. (Eds.), New Palgrave Dictionary of Economics. Palgrave Mcmillan. In press.
Diewert, W.E., Schreyer, P. (2006b). "Does capitalized net product equal discounted optimal consumption in discrete time?". Working paper. Department of Economics, University of British Columbia. http://www.econ.ubc.ca/discpapers/dp0601.pdf.
Diewert, W.E., Wales, T.J. (1992). "Quadratic spline models for producer's supply and demand functions". International Economic Review 33 (3), 705–722.
Diewert, W.E., Wales, T.J. (1995). "Flexible functional forms and tests of homogeneous separability". Journal of Econometrics 67, 259–302.
Diewert, W.E., Woodland, A.D. (2004). "The gains from trade and policy reform revisited". Review of International Economics 12 (4), 591–608.
Diewert, W.E., Wykoff, F.C. (2007). "Depreciation, deterioration and obsolescence when there is embodied or disembodied technical change". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 9: Capital and Income. Trafford Press. Chapter 6. http://www.econ.ubc.ca/diewert/dp0602.pdf.
Divisia, F. (1926). L'indice monetaire et la theorie de la monnaie. Societe anonyme du Recueil Sirey, Paris.
Domar, E.D. (1961). "On the measurement of technological change". Economic Journal L XXI, 709–729.
Dufour, A., Nakamura, A.O., Tang, J. (2007). "Productivity, business practices and high tech in the Canadian manufacturing sector". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 8: ICT and Business Process Effects. Trafford Press. Chapter 4.
Duguay, P. (1994). "Empirical evidence on the strength of the monetary transmission mechanism in Canada: An aggregate approach". Journal of Monetary Economics 33 (1), 39–61.
Duguay, P. (2006). "Productivity, terms of trade, and economic adjustment". Remarks to the Canadian Association for Business Economics, Kingston, Ontario, 28 August.
Edwards, E.O., Bell, P.W. (1961). The Theory and Measurement of Business Income. University of California Press, Berkeley.
Eichhorn, W. (1976). "Fisher's tests revisited". Econometrica 44, 247–256.
Eichhorn, W., Voeller, J. (1976). Theory of the Price Index. Lecture Notes in Economics and Mathematical Systems, vol. 140. Springer-Verlag, Berlin.
Ellerman, D., Stoker, T.M., Berndt, E.R. (2001). "Sources of productivity growth in the American coal industry: 1972–1995". In: Hulten, C.R., Dean, E.R., Harper, M.J. (Eds.), New Developments in Productivity Analysis. In: National Bureau of Economic Analysis Studies in Income and Wealth, vol. 63. University of Chicago Press, Chicago. Chapter 9.
Feenstra, R.C., Hanson, G.H. (2005). "Ownership and control in outsourcing to China: Estimating the property-rights theory of the firm". Quarterly Journal of Economics 120 (2), 729–761.
Feenstra, R.C., Reinsdorf, M.B., Slaughter, M.J., Harper, M.J. (2005). "Terms of trade gains and US productivity growth". Working paper. http://mba.tuck.dartmouth.edu/pages/faculty/matthew.slaughter/High-Tech-Feenstra-Reinsdorf-Slaughter-Harper.doc.
Fisher, I. (1911). The Purchasing Power of Money. MacMillan, London.
Fisher, I. (1922). The Making of Index Numbers. Houghton Mifflin, Boston.
Fortin, P. (1996). "Presidential address: The Great Canadian slump". Canadian Journal of Economics 29, 761–787. November.
Foster, L., Haltiwanger, J., Krizan, C.J. (2001). "Aggregate productivity growth: Lessons from microeconomic evidence". In: Hulten, C.R., Dean, E.R., Harper, M.J. (Eds.), New Developments in Productivity Analysis. In: National Bureau of Economic Analysis Studies in Income and Wealth, vol. 63. University of Chicago Press, Chicago. Chapter 8.
Fox, K.J. (2007). "Returns to scale, technical progress and total factor productivity Growth in New Zealand industries". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 7: Productivity Performance. Trafford Press. Chapter 9.

Fox, K.J., Kohli, U. (1998). "GDP growth, terms-of-trade effects, and total factor productivity". Journal of International Trade and Economic Development 7, 87–110.

Fraumeni, B. (1997). "The measurement of depreciation in the US national income and product accounts". Survey of Current Business, 7–23 (July).

Frisch, R. (1930). "Necessary and sufficient conditions regarding the form of an index number which shall meet certain of Fisher's tests". American Statistical Association Journal 25, 397–406.

Funke, H., Voeller, J. (1978). "A note on the characterization of Fisher's ideal index". In: Eichhorn, W., Henn, R., Opitz, O., Shephard, R.W. (Eds.), Theory and Applications of Economic Indices. Physica-Verlag, Wurzburg, pp. 177–181.

Funke, H., Voeller, J. (1979). "Characterization of Fisher's ideal index by three reversal tests". Statistische Hefte 20, 54–60.

Gollop, F.M. (1979). "Accounting for intermediate input: The link between sectoral and aggregate measures of productivity growth". In: National Research Council, Measurement and Interpretation of Productivity. National Academy of Sciences, Washington DC, pp. 318–333.

Gollop, F.M., Jorgenson, D.W. (1980). "US productivity growth by industry, 1947–1973". In: Kendrick, J.W., Vaccura, B. (Eds.), New Developments in Productivity Measurement and Analysis. In: NBER Studies in Income and Wealth, vol. 41. University of Chicago Press, Chicago, pp. 17–136.

Gollop, F.M., Jorgenson, D.W. (1983). "Sectoral measures of labor cost for the United States, 1948–1978". In: Triplett, J.E. (Ed.), The Measurement of Labor Cost. In: NBER Studies in Income and Wealth, vol. 44. University of Chicago Press, Chicago, pp. 185–235 and 503–520.

Greenstein S.M. (1997). "From superminis to supercomputers: Estimating surplus in the computing market". In Bresnahan and Gordon (1997, pp. 329–362).

Griliches, Z. (1960). "Measuring inputs in agriculture: A critical survey". Journal of Farm Economics 42 (5), 1411–1427.

Griliches, Z. (1963). "The sources of measured productivity growth: United States agriculture, 1940–1960". Journal of Political Economy 71 (4), 331–346.

Griliches, Z. (1997). "The Simon Kuznets Memorial Lectures" delivered in October 1997 at Yale University. These were published posthumously under the title R&D, Education and Productivity: A Personal Retrospective. Harvard University Press. 2001.

Gu, W., Tang, J. (2004). "Link between innovation and productivity in Canadian manufacturing industries". Economics of Innovation and New Technology 13 (7), 71–86.

Gullickson, W. (1995). "Measurement of productivity growth in US manufacturing". Monthly Labor Review, 13–28 (July).

Gullickson, W., Harper, M.J. (1999). Production Functions, Input–Output Tables, and the Relationship between Industry and Aggregate Productivity Measures. Bureau of Labor Statistics, Washington. February.

Haig, R.M. (1921/1959). "The concept of income: Economic and legal aspects". In: Musgrave, R.A., Shoup, C.S. (Eds.), Readings in the Economics of Taxation. Richard D. Irwin, Homewood, IL, pp. 54–76. Haig's chapter was originally published in 1921.

Hall, R.E. (1990). "Invariance properties of Solow's productivity residual". In: Diamond, P. (Ed.), Growth/Productivity/Employment. MIT Press, Cambridge, MA, pp. 71–112.

Halliwell, C. (2005). "Desperately seeking data". Feature Article, Policy Research Initiative. http://policyresearch.gc.ca/page.asp?pagenm=v8n1_art_07.

Harberger, A. (1998). "A vision of the growth process". American Economic Review 88 (1), 1–32.

Harper, M.J. (2004). "Technology and the theory of vintage aggregation". Paper presented at the Conference on Hard to Measure Goods and Services in Memory of Zvi Griliches, Washington DC, September 19–20, 2003, revised November 19, 2004.

Harper, M.J., Berndt, E.R., Wood, D.O. (1989). "Rates of return and capital aggregation using alternative rental prices". In: Jorgenson, D.W., Landau, R. (Eds.), Technology and Capital Formation. MIT Press, Cambridge, MA, pp. 331–372.

Hausman, J. (2003). "Sources of bias and solutions to bias in the CPI". Journal of Economic Perspectives 17 (1), 23–44.

Hausman, J., Leibtag, E. (2006). Consumer benefits from increased competition in shopping outlets: Measuring the effect of Wal-Mart. Journal of Applied Econometrics. In press.
Hausman, J., Leibtag, E. (2007). "Wal-Mart effects' and CPI construction". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 8: ICT and Business Process Effects. Trafford Press. Chapter 8.
Hayashi, F., Nomura, K. (2005). "Can IT be Japan's Savior?". Journal of the Japanese and International Economies 19, 543–567.
Hayek, F.A.V. (1941). "Maintaining capital intact: A reply". Economica 8, 276–280.
Hicks, J.R. (1939). Value and Capital. Clarendon Press, Oxford.
Hicks, J.R. (1940). "The valuation of the social income". Economica 7, 105–140.
Hicks, J.R. (1942). "Maintaining capital intact: A further suggestion". Economica 9, 174–179.
Hicks, J.R. (1946). Value and Capital, second ed. Clarendon Press, Oxford.
Hicks, J.R. (1961). "Measurement of capital in relation to the measurement of other economic aggregates". In: Lutz, F.A., Hague, D.C. (Eds.), The Theory of Capital. Macmillan, London.
Hicks, J.R. (1973). Capital and Time: A Neo-Austrian Theory. Clarendon Press, Oxford.
Hicks, J.R. (1981). Wealth and Welfare. Harvard University Press, Cambridge, MA.
Hill, R.J. (1999a). "Comparing price levels across countries using minimum spanning trees". The Review of Economics and Statistics 81, 135–142.
Hill, R.J. (1999b). "International comparisons using spanning trees". In: Heston, A., Lipsey, R.E. (Eds.), International and Interarea Comparisons of Income, Output and Prices. In: NBER Studies in Income and Wealth, vol. 61. The University of Chicago Press, Chicago, pp. 109–120.
Hill, R.J. (2001). "Measuring inflation and growth using spanning trees". International Economic Review 42, 167–185.
Hill, R.J. (2004). "Constructing price indexes across space and time: The case of the European Union". American Economic Review, American Economic Association 94 (5), 1379–1410 (December).
Hill, R.J. (2006). "Superlative index numbers: Not all of them are super". Journal of Econometrics 127 (1), 25–43.
Hill, R.J. (2007). "Comparing price levels and inflation rates for food across cities in Australia". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 4: International Comparisons and Trade. Trafford Press. Chapter 10.
Hill, R.J., Hill, T.P. (2003). "Expectations, capital gains and income". Economic Inquiry 41, 607–619.
Hill, R.J., Timmer, M. (2004). "Standard errors as weights in multilateral price indices". Research Memorandum GD-73. Groningen Growth and Development Center.
Hill, T.P. (1993). "Price and volume measures". In: System of National Accounts 1993. Eurostat, IMF, OECD, UN and World Bank, Luxembourg, Washington, DC, Paris, New York, and Washington, DC, pp. 379–406.
Hill, T.P. (1999). "Capital stocks, capital services and depreciation". Presented at the third meeting of the Canberra Group on Capital Measurement, Washington, DC.
Hill, T.P. (2000). "Economic depreciation and the SNA". Presented at the 26th Conference of the International Association for Research on Income and Wealth, Cracow, Poland.
Hill, T.P. (2004). Consumer Price Index Manual: Theory and Practice (CPI Manual). International Labour Office, International Monetary Fund, Organisation for Economic Co-operation and Development, Eurostat, United Nations, and The World Bank. http://www.ilo.org/public/english/bureau/stat/guides/cpi/index.htm.
Hill, T.P. (2005). "Depreciation in national accounts". Canberra II Group on Capital Measurement paper, March.
Ho, M.S., Rao, S., Tang, J. (2004). "Sources of output growth in Canadian and US industries in the information age". In: Jorgenson, D.W. (Ed.), Economic Growth in Canada and the US in the Information Age. Industry Canada Research Monograph, Ottawa.
Ho, M.S., Rao, S., Tang, J. (2007). "Measuring the contribution of ICTs to economic growth in Canadian and US industries". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 8: ICT and Business Process Effects. Trafford Press. Chapter 2.
Horngren, C.T., Foster, G. (1987). Cost Accounting: A Managerial Emphasis, sixth ed. Prentice Hall.

Hotelling, H. (1925). "A general mathematical theory of depreciation". Journal of the American Statistical Association 20, 340–353.
Hulten, C.R. (1973). "Divisia index numbers". Econometrica 41, 1017–1026.
Hulten, C.R. (1978). "Growth accounting with intermediate inputs". Review of Economic Studies 45(3) (141), 511–518.
Hulten, C.R. (1986). "Productivity change, capacity utilization, and the sources of efficiency growth". Journal of Econometrics 33 (1/2), 31–50.
Hulten, C.R. (1990). "The measurement of capital". In: Berndt, E.R., Triplett, J.E. (Eds.), Fifty Years of Economic Measurement. In: Studies in Income and Wealth, vol. 54. The University of Chicago Press, Chicago, pp. 119–152.
Hulten, C.R. (1992). "Growth accounting when technical change is embodied in capital". American Economic Review 82 (4), 964–980.
Hulten, C.R. (1996). "Capital and wealth in the revised SNA". In: Kendrick, J.W. (Ed.), The New System of National Accounts. Kluwer Academic Publishers, New York, pp. 149–181.
Hulten, C.R. (2001). "Total factor productivity: A short biography". In: Hulten, C.R., Dean, E.R., Harper, M.J. (Eds.), New Developments in Productivity Analysis. In: National Bureau of Economic Analysis Studies in Income and Wealth, vol. 63. University of Chicago Press, Chicago, pp. 1–47.
Hulten, C.R., Wykoff, F.C. (1981a). "The estimation of economic depreciation using vintage asset prices". Journal of Econometrics 15, 367–396.
Hulten, C.R., Wykoff, F.C. (1981b). "The measurement of economic depreciation". In: Hulten, C.R. (Ed.), Depreciation, Inflation and the Taxation of Income from Capital. The Urban Institute Press, Washington DC, pp. 81–125.
Inklaar R. (2006). "Cyclical productivity in Europe and the United States, Evaluating the evidence on returns to scale and input utilization". Economica (OnlineEarly Articles). doi:10.1111/j.1468-0335.2006.00554.x.
Inklaar, R., O'Mahony, M., Timmer, M.P. (2005). "ICT and Europe's productivity performance; Industry-level growth account comparisons with the United States". Review of Income and Wealth 51 (4), 505–536.
Jaffe, A.B. (1986). "Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value". American Economic Review 76, 984–1001.
Jog, V., Tang, J. (2001). "Tax reforms, debt shifting and corporate tax revenues: Multinational corporations in Canada". International Tax and Public Finance 8, 5–25.
Jorgenson, D.W. (1963). "Capital theory and investment behavior". American Economic Review 53 (2), 247–259.
Jorgenson, D.W. (1980). "Accounting for capital". In: von Furstenberg, G.M. (Ed.), Capital, Efficiency and Growth. Ballinger, Cambridge, MA, pp. 251–319.
Jorgenson, D.W. (1989). "Capital as a factor of production". In: Jorgenson, D.W., Landau, R. (Eds.), Technology and Capital Formation. MIT Press, Cambridge, MA, pp. 1–35.
Jorgenson, D.W. (1995a). Productivity, Vol. 1. Harvard University Press, Cambridge, MA.
Jorgenson, D.W. (1995b). Productivity, Vol. 2. Harvard University Press, Cambridge, MA.
Jorgenson, D.W. (1996). "Empirical studies of depreciation". Economic Inquiry 34, 24–42.
Jorgenson, D.W. (2001). "Information technology and the US economy". American Economic Review 91, 1–32.
Jorgenson, D.W. (2004). Economic Growth in Canada and the United States in the Information Age. Industry Canada Research Monograph, Ottawa.
Jorgenson, D.W., Fraumeni, B.M. (1992). "The output of the education sector". In: Griliches, Z. (Ed.), Output Measurement in the Services Sector, Z. University of Chicago Press, Chicago, IL.
Jorgenson, D.W., Gollop, F.M., Fraumeni, B.M. (1987). Productivity and US Economic Growth. Harvard University Press, Cambridge, MA.
Jorgenson, D.W., Griliches, Z. (1967). "The explanation of productivity change". Review of Economic Studies 34 (3), 249–280.
Jorgenson, D.W., Ho, M., Stiroh, K.J. (2005). "Growth of US industries and investments in information technology and higher education". In: Corrado, C., Haltiwanger, J., Sichel, D. (Eds.). Measuring Capital in a New Economy. University of Chicago Press, Chicago.

Jorgenson, D.W., Landefeld, J.S. (2006). "Blueprint for expanded and integrated US National Accounts: Review, assessment, and next steps". In: Jorgenson, Landefeld and Nordhaus (2006).
Jorgenson, D.W., Landefeld, J.S., Nordhaus, W.D. (Eds.) (2006). A New Architecture for the US National Accounts. University of Chicago Press, Chicago.
Jorgenson, D.W., Lee, F.C. (2001). "Industry-level productivity and international competitiveness between Canada and the United States". Industry Canada Research Monograph, Cat. No. C21-26/1-2000 (order information available at http://strategis.gc.ca).
Jorgenson, D.W., Motohashi, K. (2005). "Growth of US industries and investments in information technology and higher education". In: Corrado, C., Haltiwanger, J., Sichel, D. (Eds.), Measurement of Capital in the New Economy. University of Chicago Press, Chicago.
Jorgenson, D.W., Nishimizu, M. (1978). "US and Japanese economic growth, 1952–1974: An international comparison". Economic Journal 88, 707–726.
Jorgenson, D.W., Nomura, K. (2005). "The industry origins of Japanese economic growth". Journal of the Japanese and International Economies 19, 482–542.
Jorgenson, D.W., Yun, K.-Y. (1986). "Tax policy and capital allocation". Scandinavian Journal of Economics 88, 355–377.
Jorgenson, D.W., Yun, K.-Y. (1990). "Tax reform and US economic growth". Journal of Political Economy 98 (5), S151–S193.
Jorgenson, D.W., Yun, K.-Y. (1991). Tax Reform and the Cost of Capital. Clarendon Press, Oxford.
Kaplan, R.S., Atkinson, A.A. (1989). Advanced Management Accounting. Prentice Hall, Englewood Cliffs, New Jersey.
Kendrick, J.W. (1973). Postwar Productivity Trends in the United States. Columbia University Press, New York.
Kendrick, J.W. (1976). The Formation and Stocks of Total Capital. National Bureau of Economic Research, New York.
Kendrick, J.W. (1977). Understanding Productivity: An Introduction to the Dynamics of Productivity Change. Johns Hopkins University Press, Baltimore, MD.
Klette, T.J., Griliches, Z. (1996). "The inconsistency of common scale estimators when output prices are unobserved and endogenous". Journal of Applied Econometrics 11, 343–361.
Kohli, U. (1978). "A gross national product function and the derived demand for imports and supply of exports". Canadian Journal of Economics 11, 167–182.
Kohli, U. (1990). "Growth accounting in the open economy: Parametric and nonparametric estimates". Journal of Economic and Social Measurement 16, 125–136.
Kohli, U. (1991). Technology, Duality and Foreign Trade: The GNP Function Approach to Modeling Imports and Exports. University of Michigan Press, Ann Arbor.
Kohli, U. (2004). "Real GDP, real domestic income and terms of trade changes". Journal of International Economics 62, 83–106.
Kohli, U. (2005). "Switzerland's growth deficit: A real problem – but only half as bad as it looks". In: Steinmann, L., Rentsch, H. (Eds.), Diagnose: Wachstumsschwäche. Verlag Neue Zürcher Zeitung, Zurich.
Kohli, U. (2007). "Terms of trade, real exchange rates, and trading gains". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 4: International Comparisons and Trade. Trafford Press. Chapter 5.
Kuroda, M. (2006). "Toward the structural reform of official statistics: Summary translation". Economic and Social Research Institute (ESRI), Cabinet Office, Government of Japan.
Kuroda, M., Nomura, K. (2003). "Forecasting baseline CO2 emissions in Japan". In: Chang, C., Mendelsohn, R., Shaw, D. (Eds.), Global Warming and the Asian Pacific. Edward Elgar, pp. 60–74.
Kuroda, M., Nomura, K. (2004). "Technological change and accumulated capital: A dynamic decomposition of Japan's growth". In: Dietzanbacher, E., Lahr, M. (Eds.), Wassily Leontief and Input–Output Economics. Cambridge University Press, pp. 256–293.
Kuznets, S.S. (1930). Secular Movements in Production and Prices. Houghton Mifflin Co., Boston.
Laspeyres, E. (1871). "Die Berechnung einer mittleren Waarenpreissteigerung". Jahrbucher fur Nationalokonomie und Statistik 16, 296–314.

Lee, F., Tang, J. (2001a). "Multifactor productivity disparity between Canadian and US manufacturing firms". Journal of Productivity Analysis 15, 115–128.

Lee, F., Tang, J. (2001b). "Industry productivity levels and international competitiveness between Canadian and US industries". In: Jorgenson, D.W., Lee, F.C. (Eds.), Industry-Level Productivity and International Competitiveness between Canada and the United States. Industry Canada Research Monograph, Ottawa.

Leibtag, E., Nakamura, A.O., Nakamura, E., Zerom, D. (2006). "Cost pass-through in the US coffee industry". Economic Research Report Number 38. US Department of Agriculture. http://www.ers.usda.gov/publications/err38/err38fm.pdf.

Levinsohn, J., Petrin, A. (1999). "When industries become more productive, do firms? Investigating productivity dynamics". Working paper W6893. National Bureau of Economic Research.

Lipsey, R.G., Carlaw, K.I. (2004). "Total factor productivity and the measurement of technological change". Canadian Journal of Economics 37 (4), 1118–1150.

Lipsey, R.G., Carlaw, K.I., Bekar, C.T. (2006). Economic Transformations: General Purpose Technologies and Long-Term Economic Growth. Oxford University Press.

Maddison, A. (1987). "Growth and slowdown in advanced capitalist economies: Techniques of quantitative assessment". Journal of Economic Literature 25 (2), 649–698.

Malinvaud, E. (1953). "Capital accumulation and the efficient allocation of resources". Econometrica 21, 233–268.

Malmquist, S. (1953). "Index numbers and indifference surfaces". Trabajos de Estadistica 4, 209–242.

Mankiw, N.G. (2001). "The inexorable and mysterious tradeoff between inflation and unemployment". Economic Journal 111 (471), 45–61.

Mann, C.L. (2007). "Prices for international services transactions issues and a framework for development". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 4: International Comparisons and Trade. Trafford Press. Chapter 3.

Marshall, D.A. (1890). Principles of Economics. Macmillan, London.

McMahon, R.C. (1995). "Cost recovery and statistics Canada". Prepared For Presentation to The Federal/Provincial Committee on Data Dissemination, December 5; also available at http://www.stats.gov.sk.ca/docs/costrec.php.

Moorsteen, R.H. (1961). "On measuring productive potential and relative efficiency". Quarterly Journal of Economics 75, 451–467.

Morrison, C.J. (1988). "Quasi-fixed inputs in US and Japanese manufacturing: A generalized Leontief restricted cost function approach". Review of Economics and Statistics 70, 275–287.

Morrison, C.J. (1992). "Unraveling the productivity growth slowdown in the United States, Canada and Japan: The effects of subequilibrium, scale economies and markups". Review of Economics and Statistics LXXIV, 381–393.

Morrison, C.J. (1999). Cost Structure and the Measurement of Economic Performance. Kluwer Academic Publishers.

Morrison, C.J., Diewert, W.E. (1990a). "New techniques in the measurement of multifactor productivity". The Journal of Productivity Analysis 1, 265–285.

Morrison, C.J., Diewert, W.E. (1990b). "Productivity growth and changes in the terms of trade in Japan and the United States". In: Hulten, C.R. (Ed.), Productivity Growth in Japan and the United States. In: NBER Studies in Income and Wealth, vol. 53. University of Chicago Press, Chicago, pp. 201–227.

Morrison, C.J., Siegel, D. (1997). "External capital factors and increasing returns in US manufacturing". Review of Economics and Statistics, 647–654.

Muellbauer, J. (1986). "The assessment: Productivity and competitiveness in British manufacturing". Oxford Review of Economic Policy 2 (3). i-xxvi.

Nadiri, M.I. (1980). "Sectoral productivity slowdown". American Economic Review 70, 349–352.

Nadiri, M.I., Nadiri, B. (1999). "Technical change, markup, divestiture, and productivity growth in the US telecommunications industry". Review of Economics and Statistics 81 (3), 488–498 (August).

Nakajima, T., Nakamura, A.O., Nakamura, E., Nakamura, M. (abbreviated in text as N4) (2007). "Technical change in a bubble economy: Japanese manufacturing firms in the 1990s". Empirica: Journal of Applied Economics and Economic Policy 34 (3), 247–271.

Nakajima, T., Nakamura, M., Yoshioka, K. (1998). "An index number method for estimating scale economies and technical progress using time-series of cross-section data: Sources of total factor productivity growth for Japanese manufacturing, 1964–1988". Japanese Economic Review 49 (3), 310–334.

Nakajima, T., Nakamura, M., Yoshioka, K. (2001). "Economic growth: Past and present". In: Nakamura, M. (Ed.), The Japanese Business and Economic System: History and Prospects for the 21st Century. Palgrave/MacMillan/St.Martin's Press, New York, pp. 13–40.

Nakamura, A.O. (1995). "New directions for UI, social assistance, and vocational education and training". Presidential Address delivered to the annual meeting of the Canadian Economics Association, University of Quebec at Montreal, 3 June 1995 Canadian Journal of Economics 29 (4).

Nakamura, A.O., Diewert, W.E. (1996). "Can Canada afford to spend less on national statistics?". Canadian Business Economics 4 (3), 33–45.

Nakamura, A.O., Diewert, W.E. (2000). "Insurance for the unemployed: Canadian reforms and their relevance for the United States". In: Bassi, L.J., Woodbury, S.A. (Eds.), Long-Term Unemployment and Reemployment Policies. In: Research in Employment Policy, vol. 2. JAI Press, Stamford, Connecticut, pp. 217–247.

Nakamura, A.O., Lipsey, R.E. (Eds.) (2006). Services Industries and the Knowledge Based Economy. University of Calgary Press, Calgary.

Nakamura, E., Steinsson, J. (2006a). "Five facts about prices: A reevaluation of menu cost models". Working paper. http://www.people.fas.harvard.edu/~nakamura/papers/fivefacts20060808.pdf.

Nakamura, E., Steinsson, J. (2006b). "Price setting in a forward looking customer market". Working paper. http://www.people.fas.harvard.edu/~nakamura/papers/habitMar21st2006.pdf.

Nomura, K. (2004). Measurement of Capital and Productivity: Japanese Economy. Keio University Press, Tokyo (in Japanese).

Nomura, K. (2005). "Toward reframing capital measurement in Japanese national accounts". KEO Discussion Paper, No. 97.

Nomura, K., Futakamiz, T. (2005) "Measuring capital in Japan – challenges and future directions". Prepared for the 2005 OECD Working Party on National Accounts, October 11–14, Paris, France.

Nordhaus, W.D. (1982). "Economic policy in the face of declining productivity growth". European Economic Review 18, 131–157.

Nordhaus, W.D. (1997). "Do real-output and real-wage measures capture reality? The history of lighting suggests not". In Bresnahan and Gordon (1997, pp. 29–66).

OECD (2005). OECD Compendium of Productivity Indicators 2005. Prepared by Agnes Cimper, Julien Dupont, Dirk Pilat, Paul Schreyer and Colin Webb.

Olley, G.S., Pakes, A. (1996). "The dynamics of productivity in the telecommunications equipment industry". Econometrica 64 (6), 1263–1297.

Oulton, N., O'Mahony, M. (1994). Productivity and Growth. Cambridge University Press, Cambridge.

Paasche, H. (1874). "Uber die Preisentwicklung der letzten Jahre nach den Hamburger Borsennotirungen". Jahrbucher fur Nationalokonomie und Statistik 23, 168–178.

Panzar, J.C. (1989). "Technological determinants of firm and industry structure". In: Schmalensee, R., Willig, R.D. (Eds.), Handbook of Industrial Organization, vol. I. Elsevier Science Publisher, pp. 3–59.

Pavcnik, N. (2002). "Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants". Review of Economic Studies 69 (1), 245–276 (January).

Pierson, N.G. (1896). "Further considerations on index-numbers". Economic Journal 6, 127–131.

Pigou, A.C. (1941). "Maintaining capital intact". Economica 8, 271–275.

Power, L. (1998). "The missing link: Technology, investment, and productivity". Review of Economics and Statistics 80 (2), 300–313.

Prescott, E.C. (1998). "Lawrence R. Klein lecture 1997 needed: A theory of total factor productivity". International Economic Review 39 (3), 525–551.

Prud'homme, M., Sanga, D., Yu, K. (2005). "A computer software price index using scanner data". Canadian Journal of Economics 38 (3), 999–1017.

Rao, D.S.P. (2007). "The country-product-dummy method: A stochastic approach to the computation of purchasing power parities in the ICP". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O.

(Eds.), Price and Productivity Measurement, Volume 4: International Comparisons and Trade. Trafford Press. Chapter 13.

Rao, S., Tang, J., Wang, W. (2007). "What factors explain the Canada–US TFP GAP?". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 7: Productivity Performance. Trafford Press. Chapter 3.

Reinsdorf, M., Diewert, W.E., Ehemann, C. (2002). "Additive decompositions for Fisher, Törnqvist and geometric mean indexes". Journal of Economic and Social Measurement 28, 51–61.

Samuelson, P.A. (1961). "The evaluation of 'Social Income': Capital formation and wealth". In: Lutz, F.A., Hague, D.C. (Eds.), The Theory of Capital. Macmillan, London, pp. 32–57.

Samuelson, P.A. (1983). Foundations of Economic Analysis, enlarged ed. Harvard Economic Studies, vol. 80. Harvard University Press, Cambridge, MA.

Schreyer, P. (2001). OECD Productivity Manual: A Guide to the Measurement of Industry-Level and Aggregate Productivity Growth. OECD, Paris; available for download in unabridged form at http://www.oecd.org/subject/growth/prod-manual.pdf.

Schreyer, P. (2005). "International comparisons of levels of capital input and productivity". Paper presented at OECD/Ivie/BBVA workshop on productivity measurement, 17–19 October 2005, Madrid; downloadable at http://www.oecd.org/dataoecd/33/1/35446773.pdf.

Schreyer, P. (2007). "Measuring multi-factor productivity when rates of return are exogenous". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 7: Productivity Performance. Trafford Press. Chapter 8.

Schultze, C.L., Mackie, C. (Eds.) (2002). At What Price? Conceptualizing and Measuring Cost-of-Living and Price Indexes. National Academy Press, Washington DC.

Shephard, R.W. (1953). Cost and Production Functions. Princeton University Press, Princeton, NJ.

Sichel, D.E. (2001). "Productivity in the communications sector: An overview". Paper presented to Workshop on Communications Output And Productivity at the Brookings Institute, 23 February. http://www.brook.edu/es/research/projects/productivity/workshops/20010223/20010223.html (accessed 15 January 2002).

Silver, M., Heravi, S. (2003). "The measurement of quality adjusted price changes". In: Feenstra, R.C., Shapiro, M.D. (Eds.), Scanner Data and Price Indexes. In: Studies in Income and Wealth, vol. 64. University of Chicago Press, Chicago, pp. 277–316.

Smith, P. (2005). "Broadening the scope of Canada's National accounts". Horizons, Policy Research Initiative 8 (1).

Solow, R.M. (1957). "Technical change and the aggregate production function". Review of Economics and Statistics 39, 312–320.

Sterling, R.R. (1975). "Relevant financial reporting in an age of price changes". The Journal of Accountancy 139, 42–51 (February).

Stiroh, K.J. (2002). "Information technology and the US productivity revival: What do the industry data say?". American Economic Review 92 (5), 1559–1576 (December).

Tang, J., Mac Leod, C. (2005). "Labour force ageing and productivity performance in Canada". Canadian Journal of Economics 39 (2), 582–603.

Tang, J., Wang, W. (2004). "Sources of aggregate labour productivity growth in Canada and the United States". Canadian Journal of Economics 37 (2), 421–444.

Tang, J., Wang, W. (2005). "Product market competition, skill shortages and productivity: Evidence from Canadian manufacturing firms". Journal of Productivity Analysis 23 (3), 317–339.

Timmer, M.P., Inklaar, R., van Ark, B. (2005). "Alternative output measurement for the US retail trade sector". Monthly Labor Review, 39–45 (July).

Timmer, M.P., van Ark, B. (2005). "IT in the European Union: A driver of productivity divergence?". Oxford Economic Papers 51 (3).

Tinbergen, J. (1942). "Zur Theorie der langfristigen Wirtschaftsentwicklung". Weltwirtschaftliches Archiv 55 (1), 511–549. English translation (On the Theory of Trend Movements). Klaassen, L.H., Koyck, L.M., Witteveen, H.J. (Eds.), Jan Tinbergen, Selected Papers. North-Holland, Amsterdam, 1959, pp. 182–221.

Törnqvist, L. (1936). "The Bank of Finland's consumption price index". Bank of Finland Monthly Bulletin 10, 1–8.

Trefler, D. (2004). "The long and short of the Canada–US free trade agreement". American Economic Review 94 (4), 870–895.

Triplett, J.E. (1990). "The theory of industrial and occupational classifications and related phenomena". In: Proceedings, Bureau of the Census 1990 Annual Research Conference. US Department of Commence, Bureau of Census, Washington, DC, pp. 9–25.

Triplett, J.E. (1991). "Perspectives on the S.I.C.: Conceptual issues in economic classification". In: Proceedings, 1991 International Conference on the Classification of Economic Activities. US Department of Commerce, Bureau of the Census, Washington, DC, pp. 24–37.

Triplett, J.E. (1996). "Depreciation in production analysis and in income and wealth accounts: Resolution of an old debate". Economic Inquiry 34, 93–115.

Triplett, J.E. (2002). Handbook on Quality Adjustment of Price Indexes for Information and Communication Technology Products. OECD, Paris. OECD Directorate for Science, Technology and Industry.

Triplett, J.E., Bosworth, B.P. (2004). Productivity in the US Services Sector: New Sources of Economic Growth. Brookings Institution Press, Washington, DC.

van Ark, B., Inklaar, R., McGuckin, R.H. (2003). "The contribution of ICT-producing and ICT-using industries to productivity growth: A comparison of Canada, Europe and the United States". International Productivity Monitor 6, 56–63 (Spring).

van der Wiel, H.P. (1999). "Sectoral labour productivity growth". Research Memorandum No. 158, CPB Netherlands Bureau for Economic Policy Analysis, The Hague, September.

von Neumann, J. (1937). "Über ein Ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes". Ergebnisse eines Mathematische Kolloquiums 8, 73–83. Translated as: "A model of general economic equilibrium". Review of Economic Studies 12 (1945-6) 1–9.

Walsh, C.M. (1901). The Measurement of General Exchange Value. Macmillan, New York.

Walsh, C.M. (1921). "The best form of index number: Discussion". Quarterly Publication of the American Statistical Association 17, 537–544 (March).

Wang, C., Basu, S. (2007). "Risk bearing, implicit financial services and specialization in the financial industry". In: Diewert, W.E., Balk, B.M., Fixler, D., Fox, K.J., Nakamura, A.O. (Eds.), Price and Productivity Measurement, Volume 3: Services. Trafford Press. Chapter 3.

Wolff, E.N. (1996). "The productivity slowdown: The culprit at last? Follow-up on Hulten and Wolff". American Economic Review 86 (5), 1239–1252.

Wolfson, M. (1999). "New goods and the measurement of real economic growth". Canadian Journal of Economics 32 (2), 447–470.

Woodland, A.D., Turunen-Red, A. (2004). "Multilateral reforms of trade and environmental policy". Review of International Economics 12 (3), 321–336.

Yoshioka, K., Nakajima, T., Nakamura, M. (1994). "Sources of total factor productivity for Japanese manufacturing industries, 1964–1988: Issues in scale economies, technical progress, industrial policies and measurement methodologies". Monograph No. 5. Keio Economic Observatory, Keio University, Tokyo.

Chapter 67

LINKING THE THEORY WITH THE DATA: THAT IS THE CORE PROBLEM OF INTERNATIONAL ECONOMICS

EDWARD E. LEAMER

John E. Anderson Graduate School of Management, University of California, Los Angeles, Box 951481, Los Angeles, CA 90095-1481, USA
e-mail: edward.leamer@anderson.ucla.edu

Contents

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06067-9

Abstract

The greatest problem for empirical analysis is how best to allow the context to affect the inferences. Econometric theory presupposes contextual "restrictions" that can be taken as given or assigned a probability distribution. These contextual inputs are rarely available. I illustrate this point with a review of the empirical work in international economics which has focused not on properties of estimators but instead how best to link the theory with the data. I argue that the two errors we should worry about are not rejecting a true null or accepting a false null but rather taking the theory too seriously and not taking the theory seriously enough.

Keywords

Heckscher–Ohlin theory, hypothesis testing, Leontief paradox, usefulness or truthfulness

JEL classification: F11, C10, C81

1. Methodological shortcomings of econometric theory

1.1. The context matters

Econometric theory gets its questions from the numerous pathologies that afflict non-experimental data. The treatment of these pathologies invariably requires significant inputs of contextual information. For example, multicollinearity is solved by omitting "irrelevant" variables. Simultaneity is solved by selecting "appropriate" instrumental variables. Interdependence is solved by context-dependent modeling that reduces the effective dimensionality of the interdependence. Self-selection is solved by assuming "appropriate" distributions and "meaningful" selection mechanisms.

Thus the answers to the questions of how to deal with pathologies are new questions about the context: Which do you think are the least relevant variables? Which variable is a surrogate for the causal intervention you are imagining? What kind of interdependence do you think is probable? What do you think about the selection mechanism and the shape of the distributions of the observed and unobserved data?

In practice, the required contextual information is hardly ever available. This is not always a deterrent to a data analysis since the contextual information can sometimes be introduced casually without the audience and even the analyst being fully aware. For example, angst over how to pick variables in regressions has led to the formulation of automated variable selection methods by, among many others, Mallows (1973), Mayer (1975), Akaike (1974), and Schwartz (1978). The Schwartz criterion is routinely invoked today, but the same result is derived in Leamer (1978, p. 112) in a way that properly emphasizes its contextual dependence. Along the same line White (1980) has proposed corrections to standard errors that are now routinely used as if they were a treatment for any kind of heteroscedasticity. "White-washing" contrasts with the traditional approach in which the analyst is expected to commit to some model of the heteroscedasticity and thus to use corrections that are context-dependent.

In contrast with choice-of-variables and heteroscedasticity, the context-dependent exclusion restrictions needed for causal inferences in simultaneous equations settings seem painfully apparent. But despair over ever having the needed contextual information usually precedes proposals for context-free inferences. When Liu (1960) questions the identifying restrictions[1] needed for causal inference, Sims (1980) and Sargent and Sims (1977) agree, and they suggest what seems like a context-free vector-autoregressive solution. But Leamer (1985) argues that it is actually impossible to deliver on the promise of context-free causal inferences from non-experimental data. Correlation is in the data; causation is in the mind of the observer. When Judea Pearl (2000)

[1] The simultaneous equations literature presumes the existence of variables that are known to enter one "structural" equation and not to enter another. If years of schooling is endogenous and dependent on ability, birth-date is not, and since birth-date can influence years of schooling, birth-date is a surrogate for a random assignment of another couple of months of schooling. Angrist and Krueger (1991) reference. Maybe not, is your reply. Being the biggest kid in the class is what matters, not those extra couple of months of schooling.

claims "causality has been mathematized" he really means that he provides methods for "deriving causal relationships from combinations of knowledge and data" where "knowledge" is his way of saying "context."

Sniping from the sidelines have been the Bayesians who think they have a diagnosis and a solution. The problem, as they see it, is that the contextual inputs have been assumed to be certain when no one can be so sure. Their solution is simple: treat those inputs probabilistically. You do not need an exact identifying restriction, it is enough that the parameter comes from a prior distribution located at zero. You do not actually have to omit a doubtful variable to solve the collinearity problem; you can assume the coefficient comes from a distribution located at zero.

Though offering a step in a useful direction, this Bayesian literature does not offer a full solution. The difference between assuming that a parameter is exactly zero, and assuming it is has 1/3rd chance of being -1, 0 or 1, leaves us still assuming contextual inputs that may not be available.

1.2. All models are false

One critical methodological issue that we need to face is how best to input the requisite amounts of ambiguous contextual information into a data analysis. There is a second basic defect of the econometric methods that we deploy. These methods are premised on the assumption that models are either true or false, but theories derive their value precisely because they distort reality. Thus our models are neither true nor false. Our models are sometimes useful for the design of interventions and sometimes misleading. Insight, understanding and persuasiveness are the goals, not truthfulness.

The mistake of acting as though a model were true or false comes from being unclear regarding the questions the model is intended to answer. The questions of economics concern the effects of interventions such as an increase in the minimum wage, a reduction in the Federal Funds rate or a reduction of a tariff. When we build economic theory and econometric theory without clear reference to the questions, we risk building tools for tasks that do not exist. We pursue logical properties of models that may or may not be relevant. We build econometric tools for testing the truthfulness of models, when the real problem is to draw the line between settings in which a theory is useful and settings in which it is misleading. Who, when testifying in front of the Federal Reserve Board about interest rate policy, is going to discuss the "truthfulness" of a dynamic over-lapping generations model? The question is only whether that model is a useful guide for the design of monetary policy.

Once we state explicitly the intervention questions that drive the discipline, we will be reluctant to pursue all the logical implications of our models, when those implications are irrelevant to the task at hand. But the theory of econometrics does not recognize the approximate nature of our economic theory. An analyst is expected instead to act as if the economic theory were a literal and complete description of reality. A symptom that the theory is taken literally is a formal econometric "test" of the "truthfulness" of the theory. These tests routinely "accept" the theory when the sample size is small, but

routinely "reject" the theory when the data sets are large enough. To put the matter pointedly, formal hypothesis testing is only an elaborate and highly stylized ritual to determine how large is the sample.

A Bayesian approach is not fundamentally different. Models are still either true or false; it is only our knowledge that is limited.

Expressed a different way, the Type I and Type II errors that we should worry about are not accepting a false hypothesis or rejecting a true hypothesis, since theories are neither true nor false. The two relevant errors in a data analysis are treating the theory too seriously and treating the theory too casually. When we treat a model as a literal description of reality, we are being foolish. When we try to do a data analysis without any model, we are being foolhardy. The art of data analysis is to fold in just the right amount of context into the stew of data, and thus to chart a careful course between the extremes of foolishness and foolhardiness.

1.3. The goal of this paper is to broaden the conversation

In this chapter, I illustrate the dangers of the extreme approaches, drawing on examples from a field whose context I know best: international economics. One of the best-known results in the field, the so-called "Leontief Paradox" is a good example of a foolhardy approach – a data analysis without a clear theoretical foundation. Leontief found what many economists regarded to be a decisive piece of evidence against the central model of international comparative advantage – the Heckscher–Ohlin model. But Leontief made no attempt to connect his calculation with any formal model. Whoops. If you do the calculation with an explicit statement of the theory, the paradox completely unravels [Leamer (1980)].

The field now has mostly abandoned foolhardy but may be edging toward foolish. We are now starting to "test" the theory and often "rejecting" it. (And when I write "we" I include myself.) But I think this is taking the literature too literally. We do not really mean the model is not true; it is only not very accurate. And we are now trying to find ways to tinker with the model to allow it to conform better with the facts.

In the writing of this paper I am asking for a broadening of the teaching and theory of econometrics to include the possibility of approximate models and to recognize that the computation of standard errors given a set of assumptions can be a technically demanding task, but that is only one of many steps in an empirical analysis. Indeed, there is nothing of substance in the empirical literature in international microeconomics that depends on a properly computed standard error, on the properties of an estimator or on the significance level of a hypothesis test. Indeed, there is little if anything that is econometric *per se*. The commitment to a model is too minimal, and the commitment to the framework is too great. This is exploratory data analysis, not confirmatory.

The reason to publish this kind of paper in a *Handbook of Econometrics* is to remind us all that a whole branch of empirical economics is able to develop without being much affected by econometric theory. Reality can therefore be very different from our econometrics classrooms. Our textbooks and our handbooks deal almost exclusively

with formal probability theorems that cover only a portion of the problem while leaving the impression that they cover 100%. This is positively harmful to our students who memorize the blueprints but have none of the wisdom needed actually to build an influential data analysis. To quote from myself, Leamer (1993, p. 437) "When some beaming, newly minted PhD proclaims in a job seminar that he or she has managed with the assistance of the latest high-powered econometric weapon to knock off an economic theory, we offer the job for sheer brilliance, but we go on about the business of being economists without otherwise being affected by the news. Most of these new PhDs eventually 'get it', but some take longer than others."

1.4. Summary

This is therefore a highly selected survey that is intended to illustrate how context matters, and how the international theory has been linked to the data. The goal is not to cover completely the empirical work in international economics, but rather to refer to examples of empirical work that help to illustrate methodological points. For a survey of the empirical work see Leamer and Levinsohn (1995).

The formal theory of international comparative advantage is described in Section 2. In Section 3, I review first the questions of international economics, questions that should drive the data analysis. Sections 4–6 describe the empirical evidence. Tests of the *truthfulness* of the theory are the subject of Section 4. These studies implicitly or explicitly take an hypothesis testing approach and therefore treat the model as if it had truth value. Studies described in Section 5 explore the *accuracy* of the model. The borderline between Sections 4 and 5 is somewhat fuzzy, since measures of the inaccuracy of a model can be used also to determine the model's truthfulness. Section 6 surveys papers that study the *usefulness* of the framework. This is a very short section. Then, in Section 7, I refer to those papers that take the framework as given, and use it as a foundation of a study of a policy intervention. Concluding remarks comprise Section 8.

2. Theory of international comparative advantage

2.1. The elements of a model of international competition

The Heckscher–Ohlin theory which is the workhorse of international economics is basically a simple linear programming problem. A country is assumed to be endowed with a vector of factors of production, $\boldsymbol{v}$, which are used to produce a vector of products, $\boldsymbol{q}$. The ratios of inputs per unit of output are assumed to be fixed, independent of scale of production and also independent of the intensity of use of other inputs. The key assumption here is scale-independence, since substitutability of inputs can be handled by expanding the list of "products", allowing, for example, textiles made with labor-intensive techniques and textiles made with capital-intensive techniques.

The vector of inputs used to produce the vector of outputs, $\boldsymbol{q}$, is then $A\boldsymbol{q}$, where A is a matrix of fixed and given input intensities. It is further assumed that there is perfect competition in the goods and factor markets, which implies that the economy acts as if it were maximizing the value of $GDP = \boldsymbol{p}'\boldsymbol{q}$, where $\boldsymbol{p}$ is the vector of product prices determined exogenously in global marketplaces. Thus the economy operates as if it were solving the following linear programming problem, either the primal problem or the dual:

Primal program

Given the prices $\boldsymbol{p}$, choose output levels $\boldsymbol{q}$ to maximize revenues but limited by the availability of inputs:

$$\begin{aligned} &\text{Maximize} \quad \boldsymbol{p}'\boldsymbol{q} \\ &\text{subject to} \quad A\boldsymbol{q} < \boldsymbol{v} \text{ and } \boldsymbol{q} \geqslant 0. \end{aligned}$$

Dual program

Given the factor supplies $\boldsymbol{v}$, choose factor prices $\boldsymbol{w}$ to minimize costs limited by profitability of the activity:

$$\begin{aligned} &\text{Minimize} \quad \boldsymbol{w}'\boldsymbol{v} \\ &\text{subject to} \quad A'\boldsymbol{w} > \boldsymbol{p} \text{ and } \boldsymbol{w} \geqslant 0. \end{aligned}$$

2.2. *Theorems: Outputs and factor prices*

An outsider at this point is probably thinking this is a pretty silly way to model a complex economic system. What really is the point? It must come as a shock that there is a huge academic literature that explores features of this linear programming model. It is even more surprising that there are some genuine intellectual returns to all this effort.

First consider the "even" case with a matrix A that is square and invertible. Then we can write down the solutions to the primal and dual problems in terms of the inverse of A:

$$\boldsymbol{q} = A^{-1}\boldsymbol{v}, \tag{1}$$

$$\boldsymbol{w} = A'^{-1}\boldsymbol{p}. \tag{2}$$

Simplify even further and assume that the number of factors and the number of goods are equal to two and thus that A is 2 by 2. Call the factors capital and labor. Now comes the first series of remarkable results. First from (2), which determines factor rewards as a function of product prices, we have:

- **Factor Price Equalization**: An increase in the labor supply has no effect on the compensation rates of either labor or capital.

This follows from the fact that (2) does not depend on factor supplies $\boldsymbol{v}$.

The elements of A are inputs per unit of output and consequently are non-negative. The content of the next two results is driven by the fact that the inverse of a 2 by 2 positive matrix has one sign on the diagonal and the opposite sign off the diagonal. Which sign pattern depends on the ordering of the capital/labor ratios in the two sectors. Thus we get the following results:

- **Rybczynski Theorem**: An increase in the labor supply causes an increase in the output of the labor-intensive good but a fall in the output of the capital-intensive good.
- **Stolper–Samuelson Theorem**: A tax, which increases the price of the labor-intensive good, increases the compensation of labor but lowers the compensation of capital.

Still, an outsider is likely to inquire: "So what?" My answer is given in the section below titled "But what are the questions?" For now, let us complete the theory by considering other cases.

If the numbers of factors equals the number of goods but exceeds two, Equations (1) and (2) still apply and the model mathematically is completely intact. However, the model changes in very important ways in terms of its' economics implications. What remains true in higher dimensions is that the inverse of a strictly positive matrix has at least one positive and at least one negative element in every row and column [Ethier (1974)]. Thus each factor has among the goods at least one "friend" and at least one "enemy", meaning that there is at least one positive relationship between a goods price and the selected factor price, and at least one negative relationship. But the algorithm for determining the sign pattern is complex and does not yield the simple insights of the 2 by 2 model. Thus I am inclined to argue that both the Rybczynski Theorem and the Stolper–Samuelson Theorem are "true" in higher dimensions, but not "useful". The math does not depend on dimensionality; the economics does.

If the model is "uneven" with the number of goods not equal to the number of factors, things change more dramatically. I will wait until it is relevant to discuss the theory of the uneven model, but here I pose only the rhetorical question: "How should we count factors of production, and how should we count the number of products?" This rather simple question constitutes a serious attack on the very core of the discipline of economics. The subject of economics presupposes that there are "markets" that organize large numbers of identical transactions and that render the identity of the sellers and the buyers completely irrelevant. In fact, most exchanges are made in the context of long-term relationships, and are not mediated by markets that match faceless buyers with faceless sellers. The very first step in any empirical enterprise is to find an answer to this aggregation question: Which transactions can be aggregated together and treated as if they were identical, as if they were coordinated by a market? Are a Samuelson and an Arrow so similar that their work-hours can be added together to form a "high-skilled labor" aggregate? Is a Hermes scarf with a "Made in China" label so similar to a Hermes scarf with a "Made in France" label that they can be added together to form a "scarves" aggregate?

It is impossible to do empirical work in economics without implicitly or explicitly solving this aggregation/heterogeneity problem. It is particularly important here to do so because the effect of policy interventions depends very much on the number of factors and the number of products.

2.3. Theorems: International trade

This model so far is only about production and not international trade. That is appropriate because the content of the general equilibrium model comes completely from the production side. To get to trade, which is the difference between production and consumption, we need to add to the model a consumption side. One way to do this is to make an assumption that "neutralizes" the consumption side and allows external trade to behave like production. Vanek (1968) has adopted the assumption that countries have identical homothetic tastes, which is an economist's elliptical way of saying that budget shares depend on prices of products but not on income levels. Thus, if all countries face the same prices, we can write the consumption vector of country i as: $c_i = s_i q_W$ where $\boldsymbol{q}_W$ is the vector of total world outputs and s_i is the share of the world's production consumed in country i. Now with the trade vector of country i defined as $T_i = \boldsymbol{q}_W - c_i$, it is straightforward to solve for the Heckscher–Ohlin–Vanek equations that describe the factors embodied in trade as a function of the domestic and global factor supplies:

$$AT_i = A(\boldsymbol{q}_W - \boldsymbol{c}_i) = A\boldsymbol{q}_W - s_i A\boldsymbol{q}_W = \boldsymbol{v}_i - s_i \boldsymbol{v}_W. \tag{3}$$

Note the formal similarity between the trade equation (3) and the output equation (1). Thus, we have been successful in finding a consumption assumption that allows trade to behave like output.

Hidden in the symbols that comprise (3) is another important insight. Products are only bundles of factor services. A country with an abundance of labor has two ways to sell the labor services abroad. The direct way is through migration of the factor. A more subtle way is to package factor services into products, and to ship the products not the factors. Either way, arbitrage is complete and eliminates the differences in wages.

3. But what are the questions?

In the midst of all this theory, it is easy to lose track of the questions, but it is very unwise to do so. The subject of international economics exists because countries reserve the right to intervene in cross-border flows of goods, people and capital in a way that they do not intervene in strictly internal transactions. The Heckscher–Ohlin model is therefore not about why countries exchange products with each other or what they exchange, though it seems to be. The real reason that we have this model is to use it to explore the consequences of interventions that limit international exchanges, namely tariffs, quotas, capital controls, and migration restrictions. For that purpose, it has some very surprising and very important implications.

First consider the impact of migration from Mexico to the United States. Pat Buchanan would have us build walls to keep the Mexicans at home. "Not to worry" is the message of the Heckscher–Ohlin model. Provided the migrant flow leaves unchanged the mix of products made in the US and the mix of products made in Mexico, the migrant inflow leaves wages in both the US and Mexico completely unaffected. That is a very surprising implication since economists are used to thinking that there is a downward sloping demand for labor. When something surprising pops out of a theory, we need to ask why. The answer is that when we are thinking about the downward sloping demand for labor, we may not be allowing for a shift of capital from capital-intensive to labor-intensive sectors. According to the model, the potential effect of a Mexican migrant inflow into Los Angeles on native wages is completely dissipated by a shift in the product mix toward labor-intensive sectors, leaving wages completely unchanged. See Leamer (1994) for the argument that it is not the volumes of imports coming from low-wage countries that matter but the prices. Wood (1994), however, sees it differently, suggesting a volume calculation allowing for differences in technologies.

This is best expressed in terms of a pair of equations that explain the determination of wages and output mix:

$$\boldsymbol{w} = f(\boldsymbol{p}, \boldsymbol{v}),$$

$$\boldsymbol{q} = g(\boldsymbol{p}, \boldsymbol{v})$$

meaning that wages w and outputs $\boldsymbol{q}$ both depend on product prices p and on factor supplies $\boldsymbol{v}$.

One very important message of this model is that the sensitivity of wages to factor supplies is less if the sensitivity of output mix to factor supplies is more.

A second very important message of this model is that it is prices of tradables p that are the principal determinants of wages. In particular, *trade volumes are entirely irrelevant*, a point that is not understood by many economists. It does not matter if imports from low-wage countries comprise less than 3% of GDP. What matters is that a part of the workforce is producing apparel and competing in the global apparel market. That small fraction of the workforce is linked with all equivalent workers in other tradables and in nontradables through factor market arbitrage – workers choosing the best jobs available. Trade volumes are important however for one of the important targets of this model: the design of commercial policy for raising wages. When trade volumes are small, commercial policy can have only correspondingly small effects on wages.

4. Evidence regarding "truthfulness" of the model

Turning now to the empirical examination of this framework, I will try to separate studies according to their apparent purpose: to test the truthfulness of the model, to determine its accuracy or to explore its usefulness. This section deals with empirical studies that test the truthfulness of the model. Some of these studies have an explicit alternative hypothesis, but others do not.

4.1. Leontief Paradox: Why did he not use a theory?

The first and by far the most influential study of the Heckscher–Ohlin model was done by Leontief (1953) who found that US imports in 1947 were more capital intensive relative to labor than US exports. This seemed so incompatible with the clear capital abundance of the US that the finding earned the name "The Leontief Paradox". The finding was very broadly viewed as a serious empirical failure of the Heckscher–Ohlin model. This empirical "paradox" sparked a search of great breadth and intensity for a theory that could explain it. Among the explanations were labor skills, trade barriers, natural resource abundance, capital-biased consumption, and technological differences.

But what theory was Leontief really testing? Leamer (1980) showed that a carefully formulated but entirely standard Heckscher–Ohlin model allows a capital abundant country to have imports more capital intensive than exports. Indeed if one does the calculation right, properly adjusting for the US external surplus, Leontief's 1947 trade data reveal the US to have been more abundant in capital than in labor, not the other way round. This is a good illustration of the need for a clear conceptual framework when empirical work is being carried out since in its absence substantial mistakes can be made. We may think we are seeing evidence, when it is not evidence at all.[2]

4.2. Bowen–Leamer–Sveikauskas factor content "tests" are unsettling

Starting in the 1980s, there was a much greater emphasis on linking an explicit theoretical model with the data. The Heckscher–Ohlin model describes relationships between three separately observable phenomena: trade, resource supplies and technological input coefficients. A full "test" of the theory accordingly must begin with separate measures of all three of these concepts and must explore the extent to which the observed data violate the H–O restrictions. Bowen, Leamer and Sveikauskas (1987) use measurements of all three concepts and link their work carefully to a fully formulated model, namely the H–O–V model summarized by (3), which determines the factor content of trade as a function of resource supplies and external imbalance. Recognizing the impossibility of testing a theory without an alternative, these authors generalize the H–O–V model to allow (a) non-homothetic tastes characterized by linear Engel curves, (b) technological

[2] Leontief's error can be illustrated by an example with three factors and three goods. Suppose there is a capital abundant country (US), a capital scarce country (Japan), a very capital-intensive manufacture (machinery), a very labor intensive manufacture (apparel) and a moderately capital intensive agricultural activity:

$$\left(\frac{K}{L}\right)_{\text{apparel}} < \left(\frac{K}{L}\right)_{\text{agriculture}} < \left(\frac{K}{L}\right)_{\text{JAPAN}} < \left(\frac{K}{L}\right)_{\text{US}} < \left(\frac{K}{L}\right)_{\text{machinery}}.$$

Suppose further that the US has lots of land but Japan has none. The US will therefore export the agricultural good to Japan. Then, after extracting the capital used in agriculture, what the US has left for manufactures ends up with a capital abundance ratio which is less than Japan. Thus Japan then exports the most capital intensive manufacture but imports the labor-intensive manufacture. The US, which is the capital-abundant country, imports the most capital-intensive good and exports the other two.

differences among countries that affect all technological coefficients proportionately and (c) various kinds of measurement errors. In the words of Bowen, Leamer and Sveikauskas (1987, p. 805) "The data suggest errors in measurement in both trade and national factor supplies, and favor the hypothesis of neutral technological differences across countries. However, the form of the technological differences favored by the data involves a number of implausible estimates, including some in which factors yield strictly negative outputs. Thus, the Heckscher–Ohlin model does poorly, but we do not have anything that does better."

4.3. Searching for the truth

Because of the real and imagined problems empirically with the H–O model, we are now in the midst of a rather interesting search for amendments that will make the model "work". We very soon are likely to find just the right combination of amendments selected from the following five categories or possibly new ones yet to be determined:

- **Technological Differences.** Leontief's solution to the Leontief Paradox was simply to multiply the number of US workers by a productivity factor. Although the US is scarce in workers it is abundant in labor because each US worker is the equivalent of two or three or four foreign workers. This same idea has been employed unsuccessfully by BLS (1987), but more successfully by Trefler (1993), Davis and Weinstein (1998). Gabaix (1997) however sees it differently. Davis et al. (1997) also rely on technological differences to save the Heckscher–Ohlin model.
- **Home Bias**. There is not nearly so much trade as the H–O–V model suggests and not nearly so much international specialization. One reason is that distance has a strong inhibiting effect on trade, creating a strong bias in favor of local sources for most goods. Trefler (1995) has suggested one form of home bias which though theoretically suspicious works empirically quite well. Leamer (1997a) allows closeness to markets to be a source of comparative advantage. Davis and Weinstein (1998) add the gravity model to the H–O–V structure.
- **Multiple Cones.** The linear programming model allows countries to produce different subsets of products. For example, labor abundant countries may have a labor-abundant mix of products while capital-abundant countries have a capital-intensive mix, possible with some products in common. The word "cone" is a reference to the subset of factor supply vectors that all select the same mix of tradables. The H–O–V model assumes that there is only one cone, in other words that all countries have the same mix of tradables. Global general equilibrium with multiple cones of diversification is ever so much more complex than the simple model with a single cone, and much more interesting for that matter. More on cones below in the section on accuracy. Work in search of cones includes Schott (1998), Davis and Weinstein (1998), and Bernard, Jensen and Schott (2001).
- **More factors of production than tradables.** If there are enough tradables, then factor prices are set completely by external competitiveness conditions, but if the

number of traded products is less than the number of factors, then the system of zero profit conditions, $Aw = p$, cannot be solved uniquely for the factor prices. The missing markets for nontradables have to be added to the model to determine all the factor prices. These missing markets have endogenously determined prices, and the linear programming problem of maximizing the value of GDP given product prices no longer applies, since the objective function becomes a nonlinear function of the output levels. An important aspect of the solution to this nonlinear maximization model is that the internal margin affects wages. Wages depend on the demand for nontradables.

- **Preference Differences.** The Vanek assumption that budget shares are independent of incomes works well theoretically but rather badly violates the facts.

5. Evidence regarding the "accuracy" of the model

The studies referenced in the previous section act as if the problem were to decide if the model is true or false. These studies use likelihood ratios, or the equivalent, to measure the inaccuracy of the model. The studies referenced in this section use informal measures of the inaccuracy of the model, and come to no formal conclusion such as "the model can be rejected".

The large amount of intra-industry trade and the expansion of trade among the advanced developed countries while they were apparently becoming more similar are both observations that are hard to square with the Heckscher–Ohlin framework.

5.1. Intra-industry trade and trade volumes

The vast amount of "intra-industry" or "two-way" trade is an uncomfortable fact of life for the Heckscher–Ohlin framework. For example, Mercedes automobiles are shipped one way across the Channel and Jaguars go the other way. Can the H–O model help us understand the impact of British tariffs on Mercedes automobiles?

One reaction is to dismiss intra-industry trade as an artifact of the aggregation. After all, at the highest level of aggregation, with only one aggregate, there are clearly both imports and exports; indeed, if trade is balanced, exports and imports are exactly equal. Empirically, the greater is the degree of disaggregation the less is the amount of intraindustry trade. Moreover, Davis (1997) shows that the "anomaly" of intense North–North bilateral trade is compatible with the traditional Heckscher–Ohlin model. [See Davis (1995) for an amended Heckscher–Ohlin theory that can account for intra-industry trade.]

Another reaction is to patch the theory to allow it to explain intra-industry trade. For example, two-way trade in similar products is not mathematically incompatible with the Heckscher–Ohlin theory if products are allowed to differ by location of production. According to this way of thinking, the imposition of a tariff on Mercedes has a smaller

quantitative effect because Jaguars and Mercedes are imperfect substitutes, but the same qualitative effect since it raises the price of Jaguars in the UK.

Both of these reactions seem to me to represent an unnecessary clinging to the H–O model as if a single instance in which the model was not very useful meant that there are no instances in which it is useful. Better just to admit that competition between Mercedes and Jaguars is not the domain of the model. The model is relevant for the exchange between the US and Mexico of apparel and producer durables.

5.2. Cross-commodity comparisons

The Heckscher–Ohlin model has often been studied empirically with cross-commodity comparisons of export performance with industry characteristics. Early examples are Keesing (1966) and Baldwin (1971). These studies are apparently based on the assumption that the H–O model implies that some measure of export performance is positively correlated with some measure of industry capital intensity if the country is abundant in capital. I have complained about the lack of a clear theoretical foundation for many of these studies in Leamer and Levinsohn (1995), and do not need to repeat the arguments here.

5.3. Cross-country comparisons

Cross-country comparisons are another way to study the accuracy of the Heckscher–Ohlin Theorem. Studies of this type hold fixed the commodity and use the country as the experimental unit. Normally the tool of analysis is multiple regression with some measure of trade performance as the dependent variable and various characteristics of countries as the explanatory variable. Chenery (1960), Chenery and Taylor (1968) and Chenery and Syrquin (1975) were some of the earliest studies of this type although these studies did not deal with details of the structure of trade but rather with more aggregate features of the economy like the ratio of gross imports to GNP. The theory underlying many of these early cross sections regressions was casual at best, but this has given way to a formal H–O–V framework used explicitly by Leamer (1984) and more recently by Harrigan (1995, 1997) who has done careful studies of OECD production patterns as a function of their factor supplies.

5.4. Search for cones

A multiple cone model that gets the message across as simply and clearly as possible has three goods and two factors of production (capital and labor). The linear programming solution selects two out of the three goods: Countries that are abundant in capital produce the two capital-intensive goods. Countries abundant in labor produce the two labor-intensive goods. As a result, the output levels are a piecewise linear function of the factor supplies. Within each cone, the solution is linear and takes the form of (1) but the change in the mix of products from cone to cone alters the linear function. Here a

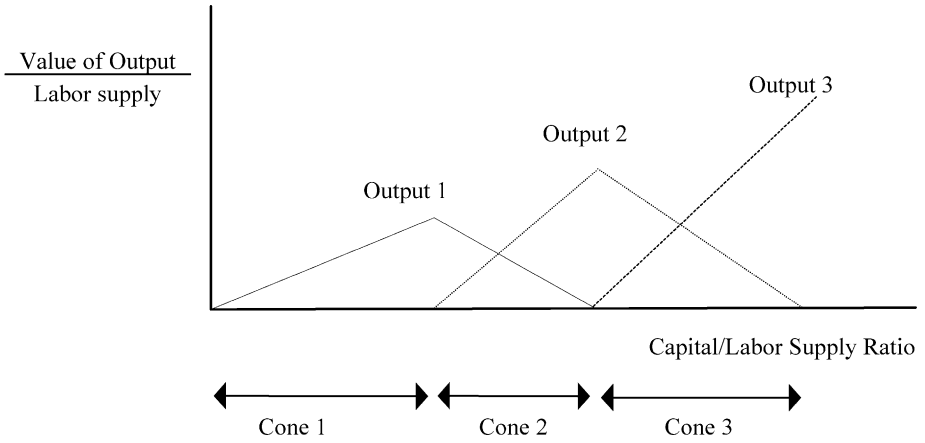

Figure 1. Cones of diversification: Three goods and two factors.

picture that describes the changing product mix in a multi-cone model with two inputs: capital and labor.

This linear programming problem underlying Figure 1 has three distinct kinds of solutions. When capital is very scarce, only good one is produced (and labor is a redundant factor with a zero wage). Accumulation of capital induces an increase in output of good one until the capital/labor supply ratio is equal to the capital/labor ratio of good 1. Then the country moves into cone 2 with goods one and two both produced. Further capital accumulation causes a reduction in the value of output of the first, most labor-intensive good, and an increase in output of good 2, which reaches a maximum when the factor supply ratio is exactly equal to capital/labor ratio of good 2. Then, with further capital accumulation, the country moves into cone 3 and starts to produce good 3.

This is an important change in the model because the effects of policy depend on the cone. Are there, for example, a North American cone (producer durables and wheat), a European cone (chemicals and machinery), a South American cone (wheat, beef and copper) and an Asian cone (apparel and footwear)? If so, trade policy advice should be different in each region.

The search for cones to date has been indecisive. Leamer (1984, 1987) and Leamer et al. (1999) look for the kind of nonlinearities suggested by Figure 1 simply by estimating a quadratic, never mind that the theory suggests that the function is piecewise linear. Leamer (1995) presents the data in Figure 2 below which plots net exports of a labor-intensive aggregate composed mostly of apparel and footwear divided by the country's workforce against the country's overall capital/labor ratio. There is very clear evidence of the nonlinearity here – countries which are very scarce in capital do not engage in much trade in these products. Exports start to emerge when the capital/labor abundance ratio is around $5000 per worker. Exports rise to around $1000 per worker when the country's abundance ratio is around $12,000 per worker. Thereafter, net ex-

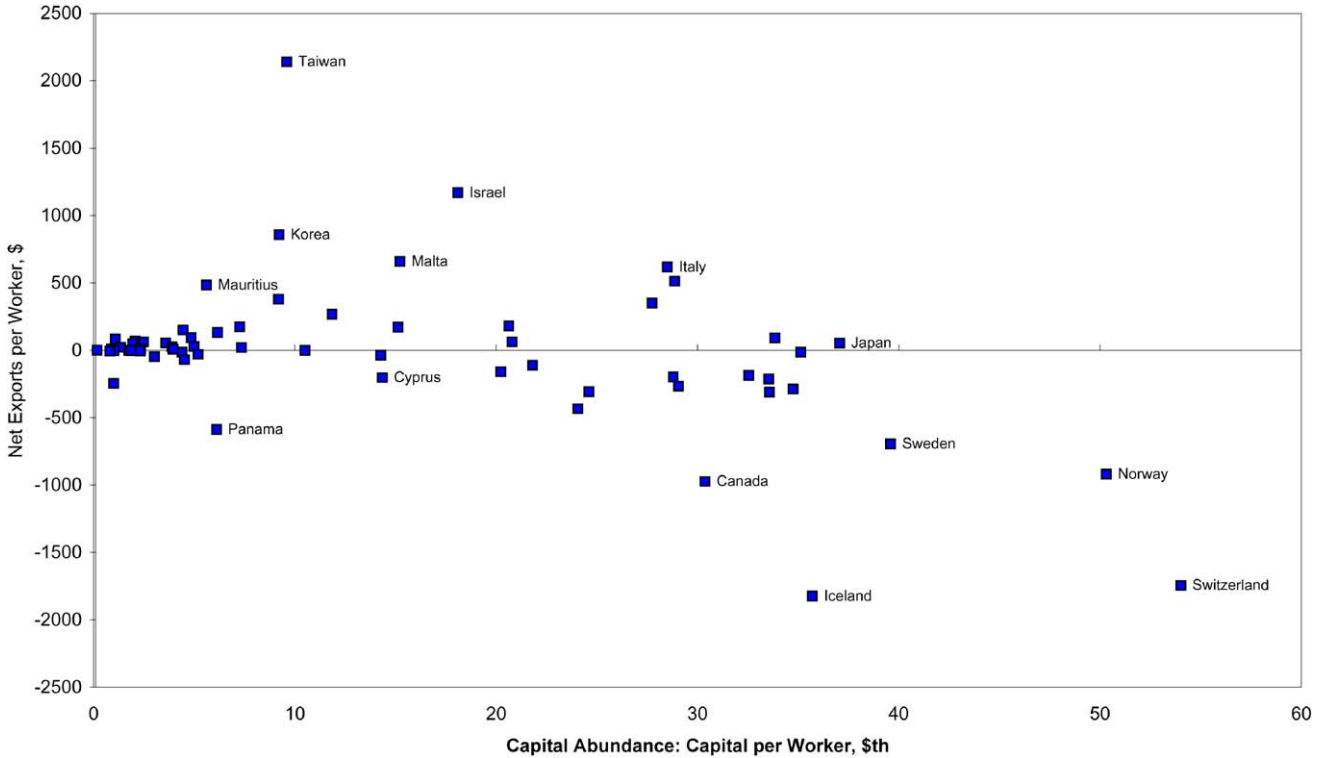

Figure 2. Net exports of labor-intensive manufactures per worker v. capital per worker, 1988.

ports steadily decline, turning negative when the country's capital/labor abundance ratio is around $20,000.

Using more disaggregated data, and taking the suggestion of a piecewise linear functional form more seriously, Schott (1998) in his dissertation work has estimated linear splines with the constraints that all the knots occur at the same points across commodities. For intriguing evidence of cones inside the US, see Bernard, Jensen and Schott (2001).

6. Evidence regarding "usefulness" of the model

The previous sections have summarized studies of the truthfulness and the accuracy of the Heckscher–Ohlin framework. A problem common to all these studies is that the use to which a model might be put is at best implicit. A model may be found to be false, or even inaccurate, and may nonetheless be useful. This section summarizes all the empirical literature on the only relevant question: How *useful* is the Heckscher–Ohlin model? Is there empirical support for the idea that this model is useful for designing trade interventions?

7. Applications of the model: The US labor market

In the 1990s the H–O model has been used as a basis for estimating the impact that trade is having on US labor markets. Again getting the theory right has been the first and most important step in the process. The H–O model has a very clear and very simple way of linking events like the liberalization of China to the US labor market. The expansion of Chinese exports of apparel drives down the global apparel price. In order to break even, apparel producers everywhere in the globe have to find some way to lower costs. If the technology is fixed, lower costs can come only from lower input prices – wages of unskilled workers for example. These lower input prices create profit opportunities in other sectors, which have to be eliminated with suitable increases in other input prices. The full solution is a mapping of product prices into factor prices $w = A'^{-1}p$.

The equation that provides the accounting framework for determining the impact of price changes on factor costs is found by differentiating a zero profit condition applicable in sector i:

$$\hat{p}_i = \boldsymbol{\theta}_i' \hat{\boldsymbol{w}} - \widehat{\mathrm{TFP}}_i, \tag{4}$$

where the carats indicate percentage changes, the subscript i the sector, p the price level, θ the vector of factor shares, w the vector of factor prices and $\widehat{\mathrm{TFP}}$ is the growth in "total factor productivity".

A very surprising feature of Equation (4) is that only the overall TFP growth in the sector matters, not the factor bias of that technological change. You are probably

thinking that there must be something wrong with any model that does not allow labor saving technological improvements to affect wages. Not to worry. There is a route by which factor bias can affect compensation levels even with Equation (4) applicable. Labor-saving technological improvements can release workers who find jobs only if the labor-intensive sectors expand, which new output can be sold only at lower prices for these labor-intensive products. In other words, *factor-biased* technological change can beget *sector-biased* product price changes which then requires changes in compensation rates to make Equation (4) hold. Thus the factor bias can matter. But be a little careful here. There is a critical intermediary in transmitting the news of factor-bias technological change to the labor markets – it is sector biased product price changes. This is exactly the same messenger that carries the news of the liberalization of China to the US labor markets. The difficulties that many of us have had in finding sector biased changes in product prices in the 1980s thus casts equivalent amounts of doubt on technological change as on globalization as an explanation of increasing wage inequality.

This leads into a more general point. In order to determine the effect of technological change on wages, we need to be very clear about the effect that technological change has on product prices. For example, if TFP growth is exactly matched by price reductions, then Equation (4) would be satisfied with no change in nominal factor earnings. The overall reduction in prices would then raise real compensation of all factors by the same percentage.

The problem of how properly to account for price changes induced by technological change is very great and entirely unresolved. The resolution requires a full general equilibrium system with a carefully specified demand side. Until this problem is resolved, we really will not have much idea of the impact of technological change on the labor market. Lawrence and Slaughter (1993) study the Stolper–Samuelson mapping from prices to wages without explicitly referring to technology. Leamer (1997b) explores a variety of ad hoc "pass-through" assumptions that allow technological improvements to be partially or fully passed on to consumers in the form of lower prices. These yield interesting results, but are not necessarily compelling.

8. Conclusions: Questions, theory and data

The academic subject of international economics has been dominated since its inception by theorists. In the 1950s, 1960s and 1970s it was Samuelson, Jones, Johnson, Bhagwati and Chipman who worked out the details of the Heckscher–Ohlin model. In 1980s it was the three men – Grossman, Helpman and Krugman – plus Dixit, Brander and Spencer who gave us a whole new set of wonderful "toys" for our theoretical playpen – models with increasing returns and imperfect competition. In the 1990s we are struggling to provide some empirical backbone for all of this theory. Though the models of the 1980s still remain in the playpen, in the sense that they resist attempts to make them genuinely empirical, the model of the 1950s, 1960s and 1970s – the Heckscher–Ohlin framework – is being subjected to a considerable amount of empirical scrutiny. We are learning

what is important about the Heckscher–Ohlin framework, and what is unimportant. We are finding out what works empirically and what does not. We are discovering ways to mend the framework but retain its basic messages. We are making progress. But not because of econometric theory. Not a single bit of our progress depends meaningfully on a standard error or a formal property of any estimator. These would require a much greater commitment to a model than most of us are willing to make. Our problem is how to connect a rich and complex theoretical structure with empirical data.

What we have learned is important methodologically: The type I and type II errors that are really important are: taking the theory too seriously and not taking the theory seriously enough. Every empirical enterprise makes some of both. The art of data analysis is to optimally trade off the two errors, using just the right level of commitment to a theory, but not too much.

References

Akaike, H. (1974). "A new look at statistical model identification". IEEE Transactions on Automatic Control AC-19, 716–723.

Angrist, J., Krueger, A.B. (1991). "Does compulsory school attendance affect schooling and earnings?". Quarterly Journal of Economics 106 (4), 979–1014 (November).

Baldwin, R.E. (1971). "Determinants of the commodity structure of US trade". The American Economic Review 61 (1), 126–146 (March).

Bernard, A., Jensen, J.B., Schott, P.K. (2001). "Factor price equality and the economies of the United States". NBER Working Paper W8068.

Bowen, H.P., Leamer, E.E., Sveikauskas, L. (1987). "Multicountry, multifactor tests of the factor abundance theory". American Economic Review 77 (5), 791–809.

Chenery, H.B. (1960). "Patterns of industrial growth". The American Economic Review 50 (4), 624–654 (September).

Chenery, H.B., Syrquin, M. (1975). Patterns of Development, 1950–1970. Oxford University Press, London.

Chenery, H.B., Taylor, L. (1968). "Development patterns: Among countries and over time". The Review of Economics and Statistics 50 (4), 391–416 (November).

Davis, D.R. (1995). "Intra-industry trade: A Heckscher–Ohlin–Ricardo approach". Journal of International Economics, 201–226 (November).

Davis, D.R. (1997). "Critical evidence on comparative advantage? North–north trade in multilateral world". Journal of Political Economy 105 (5), 1051–1060.

Davis, D.R., Weinstein, D.E. (1998). "An account of global factor trade". Mimeo. Harvard University.

Davis, D.R., Weinstein, D.E., Bradford, S.C., Shimpo, K. (1997). "Using international and Japanese regional data to determine when the factor abundance theory of trade works". American Economic Review 87 (3), 421–446.

Ethier, W. (1974). "Some of the theorems of international trade with many goods and factors". Journal of International Economics 4, 199–206.

Gabaix, X. (1997). "The factor content of trade: A rejection of the Heckscher–Ohlin Leontief hypothesis". Mimeo. Harvard University.

Harrigan, J. (1995). "Factor endowments and the international location of production: Econometric evidence for the OECD, 1970–1985". Journal of International Economics 39, 123–141.

Harrigan, J. (1997). "Technology, factor supplies and international specialization: Estimating the neoclassical model". American Economic Review 87 (4), 475–494.

Keesing, D.B. (1966). "Labor skills and comparative advantage". The American Economic Review 56 (1/2), 249–258 (March).

Lawrence, R., Slaughter, M.J. (1993). "International trade and American wages in the 1980s: Giant sucking sound or small hiccup?". Brookings Papers on Economic Activity. Microeconomics 1993 (2), 161–226.

Leamer, E.E. (1978). Specification Searches: Ad Hoc Inference with Non-Experimental Data. John Wiley and Sons, New York, NY.

Leamer, E.E. (1980). "The Leontief paradox, reconsidered". Journal of Political Economy, 495–503 (June); reprinted in Bhagwati, J. (Ed.), International Trade: Selected Readings. MIT Press, Cambridge, 1986; reprinted in Neary, J.P. (Ed.), International Trade: The International Library of Critical Writings in Economics, Edward Elgar, 1994, in press; reprinted in Kruz, H.D., Lager, C. (Eds.), Input–Output Analysis. Edward Elgar Publishing Limited, 1997.

Leamer, E.E. (1984). Sources of International Comparative Advantage. MIT Press, Cambridge, MA.

Leamer, E.E. (1985). "Vector autoregressions for causal inference?". In: Brunner, K., Meltzer, A.H. (Eds.), Carnegie–Rochester Conference Series On Public Policy: Understanding Monetary Regimes, vol. 22, pp. 255–304; reprinted in Hoover, K.D. (Ed.), The New Classical Macroeconomics. The University of California Press at Davis, 1992; reprinted in Poirier, D.J. (Ed.), The Methodology of Econometrics II. Edward Elgar, Hants, England, 1994.

Leamer, E.E. (1987). "Paths of development in the three-factor, n-good general equilibrium model". Journal of Political Economy 95 (5), 961–999.

Leamer, E.E. (1993). "Wage effects of a US–Mexican free trade agreement". In: Garber, P.M. (Ed.), The Mexico–US Free Trade Agreement. MIT Press, Cambridge, MA, pp. 57–125.

Leamer, E.E. (1994). "Trade, wages and revolving door ideas". NBER Working Paper No. 4716. April.

Leamer, E.E. (1995). "The Heckscher–Ohlin model in theory and practice". Princeton Studies in International Finance 77, 1–45.

Leamer, E.E. (1997a). "Access to western markets and eastern effort". In: Zecchini, S. (Ed.), Lessons from the Economic Transition, Central and Eastern Europe in the 1990s. Kluwer Academic Publishers, Dordrecht, pp. 503–526.

Leamer, E.E. (1997b). "In search of Stolper–Samuelson effects on US wages". In: Collins, S. (Ed.), Imports, Exports and the American Worker. Brookings.

Leamer, E.E., Levinsohn, J. (1995). "International trade theory: The evidence". In: Grossman, G., Rogoff, K. (Eds.), Handbook of International Economics, vol. III, pp. 1339–1394.

Leamer, E.E., Maul, H., Rodriguez, S., Schott, P.K. (1999). "Does natural resource abundance increase Latin American income inequality?". Journal of Development Economics 59 (1), 3–42.

Leontief, W. (1953). "Domestic production and foreign trade; The American capital position re-examined". Proceedings of the American Philosophical Society 97 (4), 332–349 (September).

Liu, T.C. (1960). "Underidentification, structural estimation, and forecasting". Econometrica 28, 855–865.

Mallows, C.L. (1973). "Some comments on Cp". Technometrics 15, 661–675.

Mayer, T. (1975). "Selecting economic hypotheses by goodness of fit". Economics Journal 85, 877–882.

Pearl, J. (2000). Causality: Models, Reasoning and Inference. Cambridge University Press, Cambridge.

Sargent, T., Sims, C. (1977). "Business cycle modelling without pretending to have too much a priori economic theory". In: Sims, C. (Ed.), New Methods of Business Cycle Research. Federal Reserve Bank of Minneapolis, Minneapolis.

Schwartz, G. (1978). "Estimating the dimension of a model". Annals of Statistics 6, 461–464.

Schott, P. (1998). "One size fits all?". Chapter one of PhD dissertation, 1999.

Sims, C. (1980). "Macroeconomics and reality". Econometrica 48, 1–48.

Trefler, D. (1993). "International factor price differences – Leontief was right!". Journal of Political Economy 101 (6), 961–987.

Trefler, D. (1995). "The case of the missing trade and other mysteries". American Economic Review 85 (5), 1029–1046.

Vanek, J. (1968). "The factor proportions theory: The N-factor case". Kyklos 21 (4), 749–756 (October).

Wood, A. (1994). North–South Trade, Employment and Inequality. Clarendon Press, Oxford.

White, H. (1980). "A heteroscedasticity-consistent covariance matrix estimator and a direct test of heteroscedasticity". Econometrica 48, 817–838.

Chapter 68

MODELS OF AGGREGATE ECONOMIC RELATIONSHIPS THAT ACCOUNT FOR HETEROGENEITY*

RICHARD BLUNDELL

Institute for Fiscal Studies and Department of Economics, University College London, Gower Street, London, WC1E 6BT, UK

THOMAS M. STOKER

Sloan School of Management, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02138, USA

Contents

* A previous version of this chapter was published as "Heterogeneity and Aggregation" in the *Journal of Economic Literature*, vol. 43, 2, pp. 347–391. We are grateful for comments from numerous colleagues, especially Orazio Attanasio, Martin Browning, Jim Heckman, Arthur Lewbel and the reviewers. Thanks are also due to Zoë Oldfield and Howard Reed who helped organise the data for us. Financial support from the ESRC Centre for the Microeconomic Analysis of Fiscal Policy at IFS is gratefully acknowledged. The usual disclaimer applies.

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06068-0

Abstract

This chapter covers recent solutions to aggregation problems in three application areas: consumer demand analysis, consumption growth and wealth, and labor participation and wages. Each area involves treatment of heterogeneity and nonlinearity at the individual level. Three types of heterogeneity are highlighted: heterogeneity in individual tastes, heterogeneity in income and wealth risks, and heterogeneity in market participation. Work in each area is illustrated using results from empirical data. The overall aim is to present specific models that connect individual behavior with aggregate statistics, as well as to discuss the principles for constructing such models.

Keywords

aggregation, heterogeneity, consumption growth, consumer demand, wage growth, sample selection

JEL classification: D1, D3, D91, E21, J31, C51

1. Introduction

Models of optimal behavior typically apply at the individual level. Important issues of economic policy typically apply to large groups or entire economies. If different individuals behaved in essentially the same way, then group statistics would mirror that common behavior. However, this is never the case. In virtually every application area, there is evidence of extensive differences in behavior across individuals, or individual heterogeneity. This is true whether the "individual agent" is a single person, a household, a firm or some other decision-making entity that is relevant for the applied question of interest. Individual differences are a fact of life. They create a paramount difference between behavior at the individual level and aggregate statistics for an economy or large group. Resolving that difference involves solving the problem of aggregation.

Some broad properties will transmit between the individual level and aggregate statistics, but typically not enough for quantitative analysis. For instance, consider spending behavior by households when the price of a good increases. If some households decrease their purchases of that good and no household increases its purchases, then aggregate spending on that good must decrease. However, knowing that aggregate spending will decrease is very different from knowing the exact amount it will decrease, or how sensitive aggregate spending is to that price increase. This latter information is required for any analysis of aggregate demand–supply impacts or impacts of policy such as tax changes. To pin down a value to use as the "aggregate price elasticity", one must come to grips with how individual households differ across the population. In an econometric setting, this requires explicit modeling of household behavior and the differences relevant to spending. Moreover, this example involves the simplest type of change, namely a common price change faced by all households. The issues multiply when the changes themselves vary across households. Consider spending impacts that arise from a policy that injects new income into the household sector. Now it is not even clear how to make sense of a value of an "income elasticity" for the economy. If the policy targets poor households, then one likely gets different impacts than if the policy targeted middle-income households.

The issues raised by aggregation are not new, but rather have been a part of the discussion of empirical work in economics for much of the past century. What is new is the development of econometric models and methods that explicitly deal with aggregation problems. Such models apply equally well to individual data and to aggregate level statistics. The purpose of this chapter is to cover these developments in a few selected application areas. We include discussion of the principles that guide the construction of such econometric models.

In line with the models we cover, we take a practical approach to the definitions of "individual level" and "aggregate statistics", sidestepping a number of issues about the underpinnings of standard econometric models. For instance, empirical practice in demand modeling is to define a household as the "individual agent" and study total (economy-wide) expenditures on various commodities as the aggregate statistics of in-

terest. Thus, at the micro level, purchase decisions made by individual family members are assumed to be in line with an overall household plan, which sidesteps the issues raised by bargaining within a family. In terms of the macro level, modeling total or "per capita" commodity expenditures entails different issues from modeling an alternative type of aggregate statistic. For instance, different issues would arise if one wanted to study what fraction of households fell below a minimum threshold for food expenditures, or if one wanted to study inequality in living standards from the Gini coefficient of food expenditures.

While practical, this posture is not arbitrary. Much of the development we discuss has been made possible by the enhanced availability of survey data on individual behavior. Standard microeconometric models (and the assumptions that justify them) provide the most natural starting point for building models that account for aggregation. The appropriate choice of aggregate statistics is driven by data availability and the policy question of interest. Economy-wide totals or averages (from national accounts) are the most commonly available statistics for modeling. They are also the most important statistics for most questions of economic policy, such as questions involving prices, interest rates, total savings, market demand and supply, total tax revenues, aggregate wages and unemployment.

Given the application area, an econometric model that accounts for aggregation consists of individual-level equations and equations for aggregate statistics. Ideally, the individual equations will capture all important economic effects and allow for realistic individual heterogeneity. The aggregate equations must be fully consistent with the individual equations, and typically will require assumptions on the distribution of individual heterogeneity. Taken as a whole, these equations constitute a single model that relates to data at all levels – individual cross-section or panel data and aggregate statistics over time. All relevant data sources can be used in estimation, and an estimated model can be applied to any level – individual, proper subgroup or the full economy. There are multiple levels of testable implications of such a model: from the individual model, from the aggregate equations and from the necessary distributional assumptions.

This chapter covers specific models and related work in three application areas: consumer demand analysis, consumption and saving analysis and analysis of wages and labor market participation. A key issue is to identify what kinds of individual differences, or heterogeneity, are relevant for each application area. As an organizing principle, we consider (i) heterogeneity in individual tastes and incomes, (ii) heterogeneity in wealth and income risks faced by individuals and (iii) heterogeneity in market participation.[1] There is a generic tension between the degree of individual heterogeneity accounted for and the ease with which one can draw implications for economic aggregates. We point out how different types of heterogeneity are accommodated in the different application areas.

[1] This roughly coincides with the categorization of heterogeneity discussed in Browning, Hansen and Heckman (1999).

We are concerned with models that strike a balance between realism (flexibility), adherence to restrictions from economic theory and connections between individual behavior and aggregate statistics. We consider several settings where individual models are intrinsically nonlinear, and for those we must make specific assumptions on the distributions of relevant heterogeneous characteristics. We present results that can be used to explore the impact of heterogeneity in empirical applications, that assume reasonable (and hopefully plausible) parameterizations of both individual equations and distributions of heterogeneity. We do not go into details about estimation; for each application area, we present models with empirically plausible equations for individuals and consistent equations for the relevant economic aggregates. Again, the point is to develop a single framework that has the ability to address empirical issues at the individual (micro) level, the aggregate (macro) level or both.

We begin with our coverage of consumer demand models in Section 2, the area which has seen the most extensive development of solutions to aggregation problems. The difficult issues in consumer demand include clear evidence of nonlinearity in income effects (e.g. Engel's Law for food) and pervasive evidence of variations in demand with observable characteristics of households. We discuss each of these problems in turn, and use the discussion to cover traditional results as well as "aggregation factors" as a method of empirically studying aggregation bias. We cover recent empirical demand models, and present aggregation factors computed from data on British households. That is, we cover the standard issues faced by aggregating over heterogeneous households in a static decision-making format, and illustrate with application to empirical demand models in current use. We close with a discussion of recent work that studies aggregate demand structure without making specific behavioral assumptions on individual demands.

In Section 3 we discuss models of overall consumption growth and wealth. Here we must consider heterogeneity in tastes, but we focus on the issues that arise from heterogeneity in income shocks, showing how different types of shocks transmit to aggregate consumption. We start with a discussion of quadratic preferences in order to focus on income and wealth, and then generalize to recent empirical models that permit precautionary saving. Because of the log-linear form of these models, we must make explicit distributional assumptions to solve for aggregate equations. We cover the types of heterogeneity found in consumption relationships, as well as various other aspects of our modeling, illustrating with empirical data. We follow this with a brief discussion of modeling liquidity constraints and the impacts on aggregate consumption. We close this section with a discussion of recent progress in general equilibrium modeling of consumption, saving and wealth.

Section 4 covers recent work on labor participation and aggregate wage rates. The main issues here concern how to interpret observed variations in aggregate wages – are they due to changes in wages of individuals or to changes in the population of participating workers? We focus on the issues of heterogeneity in market participation, and develop a paradigm that allows isolation of the participation structure from the wage structure. This involves tracking the impacts of selection on the composition of

the working population, the impacts of weighting individual wage rates by hours in the construction of aggregate wages and the impact of observed wage heterogeneity. We show how accounting for these features gives a substantively different picture of the wage situation in Britain from that suggested by observed aggregate wage patterns. Here we have a situation where there is substantial heterogeneity and substantial nonlinearity, and we show how to address these issues and draw conclusions relevant to economic policy.

Section 5 concludes with some general observations on the status of work on aggregation in economics.

This chapter touches on many of the main ideas that arise in addressing aggregation problems, but it is by no means a comprehensive survey of all relevant topics or recent approaches to such problems. For instance, we limit our remarks on the basic nature of aggregation problems, or how it is senseless to ascribe behavioral interpretations to estimated relationships among aggregate data without a detailed treatment of the links between individual and aggregate levels. It is well known that convenient constructs such as a "representative agent" have, in fact, no general justification – we will not further belabor their lack of foundation. See the surveys by Stoker (1993) and Browning, Hansen and Heckman (1999) for background on these basic problems. It is useful to mention two related lines of research, that we do not cover. The first is the work on how economic theory provides few restrictions on market excess demands – see Sonnenschein (1972) and Schafer and Sonnenschein (1982) among others, and Brown and Matzkin (1996) for a more recent contribution. The second is the work on collective decision making within households as pioneered by Chiappori (1988, 1994).

We will also limit our attention to aggregation over individuals, and not discuss the voluminous literature on aggregation over commodities. This latter literature concerns the construction of aggregate "goods" from primary commodities, as well as the consistency of multistage budgeting and other simplifications of choice processes. While very important for empirical work, the issues of commodity aggregation apply within decision processes of individuals and, as such, would take us too far afield of our main themes. See the survey by Blundell (1988) as well as the book by Blackorby, Primont and Russell (1978) for background on commodity aggregation and multistage budgeting. We do not cover the growing literature on hedonic/characteristics models, which can serve to facilitate commodity aggregation or other simplifications in decision making. Moreover, we do not cover recent advances that use aggregation to solve microeconometric estimation problems: see Imbens and Lancaster (1994) for the basic approach and Berry, Levinsohn and Pakes (2004) for a recent application to estimation of demands for differentiated products.

Finally, we do not cover in great detail work that is associated with time-series aggregation. That work studies how the time-series properties of aggregate statistics relate to the time-series processes of associated data series for individuals, such as stationarity and co-integration. To permit such focus, that work relies on strictly linear models for individual agents, which again turn the discussion away from heterogeneity in individual reactions and other behavior. We do make reference to time-series properties

of income processes as relevant to our discussion of individual and aggregate consumption, but do not focus on time-series properties in any general way. Interested readers can pursue Granger (1980, 1987, 1990) and the book by Forni and Lippi (1997) for more comprehensive treatment of this literature.[2]

2. Consumer demand analysis

We begin with a discussion of aggregation and consumer demand analysis. Here the empirical problem is to characterize budget allocation to several categories of commodities. The individual level is that of a household, which is traditional in demand analysis. The economic aggregates to be modeled are average (economy-wide, per household) expenditures on the categories of commodities. We are interested in aggregate demand, or how average category expenditures relate to prices and the distribution of total budgets across the economy.

In a bit more detail, we assume that households have a two-stage planning process, where they set the total budget for the current period using a forward-looking plan, and then allocate that current budget to the categories of nondurable commodities.[3] As such, we are not concerned with heterogeneity in the risks faced by households in income and wealth levels – they have already been processed by the household in their choice of total budget (and, possibly, in their stocks of durable goods). We consider commodity categories that are sufficiently broad that household expenditures are nonzero (food categories, clothing categories, etc.), and so we are not concerned with zero responses, or heterogeneity in market participation.

We are concerned with heterogeneity in total budgets and in needs and tastes. It is a well-known empirical fact that category expenditure allocations vary nonlinearly with total budget size (for instance, Engel's Law with regard to food expenditures). Early applications of exact aggregation demand systems had budget shares in semi-log form (with or without attributes), namely the popular Translog models of Jorgenson, Lau and Stoker (1980, 1982) and Almost Ideal models of Deaton and Muellbauer (1980a, 1980b) respectively. More recent empirical studies have shown the need for further nonlinear terms in certain expenditure share equations. In particular, evidence suggests that quadratic logarithmic income terms are required [see, for example, Atkinson, Gomulka and Stern (1990), Bierens and Pott-Buter (1990), Hausman, Newey and Powell (1994), Härdle, Hildenbrand and Jerison (1991), Lewbel (1991) and Blundell, Pashardes and Weber (1993)]. This nonlinearity means that aggregate demands will be affected by total budget size as well as the degree of inequality in budgets across consumers. It is

[2] See Stoker (1986c, 1993), Lewbel (1994) and others for examples of clear problems in inferring behavioral reactions from time-series results in the presence of individual heterogeneity.

[3] Provided that intertemporal preferences are additive, this accords with a fairly general intertemporal model of expected utility maximization [see Deaton and Muellbauer (1980b), among others].

also well known that category expenditures vary substantially with demographic composition of households, such as how many children are present, or whether the head of household is young or elderly [see Barten (1964), Pollak and Wales (1981), Ray (1983) and Browning (1992)].

Our aim is to understand how behavioral effects for households impinge on price effects and distributional effects on aggregate demands. Understanding these effects is a key ingredient in understanding how the composition of the population affects demand growth over time and relative prices across the different commodity categories.

2.1. Aggregation of consumer demand relationships

Our framework requires accounting for individuals (households), goods and time periods. In each period t, individual i chooses demands q_{ijt} (or equivalently expenditures $p_{jt}q_{ijt}$) for $j = 1, \ldots, J$ goods by maximizing preferences subject to an income constraint, where $i = 1, \ldots, n_t$. Prices p_{jt} are assumed to be constant across individuals at any point in time, with $p_t = (p_{1t}, \ldots, p_{Jt})$ summarizing all prices. Individuals have total expenditure budget $m_{it} = \sum_j p_{jt}q_{ijt}$, or income for short,[4] and are described by a vector of household attributes z_{it}, such as composition and demographic characteristics. The general form for individual demands is written

$$q_{ijt} = g_{jt}(p_t, m_{it}, z_{it}). \tag{1}$$

This model reflects heterogeneity in income m_{it} and individual attributes z_{it}. Specific empirical models involve the specification of these elements,[5] including a parametric formula for g_{jt}.

Economy-wide average demands and average income are

$$\frac{\sum_i q_{ijt}}{n_t}, \quad j = 1, \ldots, J, \quad \text{and} \quad \frac{\sum_i m_{it}}{n_t}. \tag{2}$$

We assume that the population of the economy is sufficiently large to ignore sampling error, and represent these averages as the associated population means

$$E_t(q_{ijt}), \quad j = 1, \ldots, J, \quad \text{and} \quad E_t(m_{it}). \tag{3}$$

Our general framework will utilize various other aggregates, such as statistics on the distribution of consumer characteristics z_{it}.

[4] It is common parlance in the demand literature to refer to "total budget expenditure" as "income", as we do here. In the later section on consumption, we return to using "income" more correctly, as current consumption expenditures plus saving.

[5] For most of our discussion, z_{it} can be taken as observable. When we discuss explicit empirical models, we will include unobserved attributes, random disturbances, etc.

2.1.1. Various approaches: Exact aggregation and distributional restrictions

We begin by discussing various approaches to aggregation in general terms. From (1), aggregate demand is given formally as

$$E_t(q_{ijt}) = \int g_{jt}(p_t, m_{it}, z_{it})\,\mathrm{d}F_t(m_{it}, z_{it}), \tag{4}$$

where $F_t(m_{it}, z_{it})$ is the cross-section distribution of income and attributes at time t. At the simplest level, approaches to aggregation seek a straightforward relationship between average demand, average income and average attribute values

$$E_t(q_{ijt}) = G_{jt}\big(p_t, E_t(m_{it}), E_t(z_{it})\big). \tag{5}$$

The *exact aggregation* approach is based on linearity restrictions on individual preferences/demands g_{ijt} that allow the relationship G_{jt} to be derived in a particularly simple way, such that knowledge of G_{jt} is sufficient to identify (the parameters of) the individual demand model. Take, for example,

$$g_{jt}(p_t, m_{it}, z_{it}) = b_{0j}(p_t)m_{it} + b_{1j}(p_t)m_{it}\ln m_{it} + b_{2j}(p_t)m_{it}z_{it}, \tag{6}$$

where we suppose z_{it} is a single variable that has $z_{it} = 1$ for an elderly household and $z_{it} = 0$ otherwise. Individual demand has a linear term in income and a nonlinear term in income, and the slope of the linear term is different for elderly households. All of these slopes can vary with p_t. Now, aggregate demand is

$$E_t(q_{ijt}) = b_{0j}(p_t)E_t(m_{it}) + b_{1j}(p_t)E(m_{it}\ln m_{it}) + b_{2j}(p_t)E(m_{it}z_{it}), \tag{7}$$

which depends on average income $E_t(m_{it})$ and two other statistics, $E(m_{it}\ln m_{it})$ and $E(m_{it}z_{it})$. The coefficients are the same in the individual and aggregate models, which is the bridge through which individual preference parameters manifest in aggregate demands (and can be recovered using aggregate data).

In order to judge the impact of aggregation on demand, it is convenient to use *aggregation factors*.[6] Write aggregate demand as

$$\begin{aligned} E_t(q_{ijt}) = {} & b_{0j}(p_t)E_t(m_{it}) + b_{1j}(p_t)\pi_{1t}E(m_{it})\ln E(m_{it}) \\ & + b_{2j}(p_t)\pi_{2t}E(m_{it})E(z_{it}), \end{aligned} \tag{8}$$

where

$$\pi_{1t} = \frac{E(m_{it}\ln m_{it})}{E(m_{it})\ln E(m_{it})} \quad \text{and} \quad \pi_{2t} = \frac{E(m_{it}z_{it})}{E(m_{it})E(z_{it})}. \tag{9}$$

The factors π_{1t} and π_{2t} show how the coefficients in (7) are adjusted if individual demand is evaluated at average income and average attributes, as in (8). π_{1t} reflects inequality in the income distribution through the entropy term $E(m_{it}\ln m_{it})$ and π_{2t}

[6] The use of aggregation factors was first proposed by Blundell, Pashardes and Weber (1993).

reflects the distribution of income of the elderly, as the ratio of the elderly's share in aggregate income $E(m_{it}z_{it})/E(m_{it})$ to the percentage of elderly $E(z_{it})$ in the population. Aggregation factors are useful for two reasons. First, if they are stable, then aggregate demand has similar structure to individual demand. Second, their average value indicates how much bias is introduced in estimation using aggregate data alone.[7]

In contrast, the *distributional* approach considers restrictions on the heterogeneity distribution $F_t(m_{it}, z_{it})$. Suppose the density $\mathrm{d}F_t(m_{it}, z_{it})$ is an explicit function of $E_t(m_{it})$, $E(z_{it})$ and other parameters, such as variances and higher-order moments. Then with a general nonlinear specification of individual demands g_{ijt}, we could solve (4) directly, expressing aggregate demand $E_t(q_{ijt})$ as a function of those distributional parameters. Here, recovery of individual demand parameters from aggregate demand would be possible with sufficient variation in the distribution $F_t(m_{it}, z_{it})$ over t.[8]

While conceptually different from exact aggregation, the distributional approach should not be thought of as a distinct alternative in empirical modeling. With distribution restrictions, formulating a model via direct integration in (4) may be difficult in practice. As such, distributional restrictions are often used together with exact aggregation restrictions, combining simplifying regularities of the income-attribute distribution with linearity restrictions in individual demands.

One example is with mean-scaling, as discussed in Lewbel (1990), where the distribution of income does not change relative shape but just scales up or down. Mean-scaling can arise with a redistribution mechanism where individual budgets are all scaled the same, as in $m_{it} = m_{it-1}(E_t(m_{it})/E_{t-1}(m_{it-1}))$. This structure allows distributional statistics such as those in (7) to be computed from mean income only.

Another example arises from (distributional) exclusion restrictions. Certain attributes can be excluded from aggregate demand if their distribution conditional on income is stable over time; if

$$\mathrm{d}F_t(m_{it}, z_{it}) = f_z(z_{it}|m_{it})\,\mathrm{d}F_t^*(m_{it}) \tag{10}$$

where $f_z(z_{it}|m_{it})$ does not vary with t, then from (4),

$$\begin{aligned} E_t(q_{ijt}) &= \int g_{jt}(p_t, m_{it}, z_{it})\, f_z(z_{it}|m_{it})\,\mathrm{d}F_t^*(m_{it}) \\ &= \int g_{jt}^*(p_t, m_{it})\,\mathrm{d}F_t^*(m_{it}). \end{aligned} \tag{11}$$

That is, z_{it} and its distributional statistics are excluded from the equation for aggregate demand. Aggregate demand reflects heterogeneity only through variation in the income

[7] For instance, in (8), $b_{1j}(p_t)$ is the coefficient of $E(m_{it}\ln m_{it})$, whereas $b_{1j}(p_t)\pi_{1t}$ is the coefficient of $E(m_{it})\ln E(m_{it})$. If π_{1t} is stable, $\pi_{1t} = \pi_0$, then $b_{1j}(p_t)\pi_{1t}$ is proportional to $b_{1j}(p_t)$. In this sense, the structure of aggregate demand matches that of individual demand, but the use of aggregate data alone would estimate the individual coefficient with a proportional bias of π_0.

[8] Technically, what is necessary for recoverability is completeness of the class of income-attribute distributions; see Stoker (1984a).

distribution – there is not enough variation in the z_{it} distribution over t to recover the individual effects from aggregate demand. We discuss various other examples of partial distribution restrictions below.

2.1.2. Demand and budget share models

There has been a substantial amount of work on the precise structure of individual preferences and demands consistent with exact aggregation. The most well-known result of this kind is in the extreme case where the aggregate model simply relates average demands $E_t(q_{ijt})$ to the vector of relative prices p_t and average expenditure $E_t(m_{it})$. Gorman (1953) showed that this required preferences to be quasi-homothetic, with individual demands linear in m_{it}.

Omitting reference to attributes z_{it} for now, the general formulation for exact aggregation has demands of the form

$$q_{ijt} = a_{0j}(p_t) + b_{0j}(p_t)h_0(m_{it}) + \cdots + b_{Mj}(p_t)h_M(m_{it}) \tag{12}$$

with aggregate demands given as

$$E_t(q_{ijt}) = a_{0j}(p_t) + b_{0j}(p_t)E_t\big[h_0(m_{it})\big] + \cdots + b_{Mj}(p_t)E_t\big[h_M(m_{it})\big]. \tag{13}$$

As above, provided there is sufficient variation in the statistics $E_t[h_0(m_{it})], \ldots, E_t[h_M(m_{it})]$, the coefficients $a_{0j}(p_t), b_{0j}(p_t), \ldots, b_{Mj}(p_t)$, and hence individual demands, can be fully recovered from aggregate data.

Lau (1977, 1982) originally proposed the exact aggregation framework, and demonstrated that demands of the form (12) were not only sufficient but also necessary for exact aggregation, or aggregation without distributional restrictions [c.f. Stoker (1993) and Jorgenson, Lau and Stoker (1982)]. Muellbauer (1975) studied a related problem, and established results for the special case of (12) with only two income terms.[9] These both showed several implications of applying integrability restrictions to (12). If demands are zero at zero total expenditure, then $a_{0j}(p_t) = 0$. The budget constraint implies that one can set $h_0(m_{it}) = m_{it}$, without loss of generality. With homogeneity of degree zero in prices and incomes, one can assert the forms of the remaining income terms, which include the entropy form $h_1(m_{it}) = m_{it}\ln m_{it}$ and the power form $h_1(m_{it}) = m_{it}^{\theta}$. This theory provides the background requirements for specific exact aggregation demand models, such as those we discuss below.[10]

The tradition in empirical demand analysis is to focus on relative allocations, and estimate equations for budget shares. The exact aggregation form (12) is applied to budget shares for this purpose. In particular, if we set $a_{0j}(p_t) = 0$ and $h_0(m_{it}) = m_{it}$

[9] Muellbauer (1975) studied the conditions under which aggregate budget shares would depend only on a single representative income value, which turned out to be analogous to the exact aggregation problem with only two expenditure terms.

[10] See also Lewbel (1989b, 1991, 1993) and Stoker (1984a, 1984b).

in (12), then budget shares $w_{ijt} = p_{jt}q_{ijt}/m_{it}$ take on a similar linear form. We have

$$w_{ijt} = \frac{p_{jt}q_{ijt}}{m_{it}} = b_{0j}(p_t) + b_{1j}(p_t)h_1(m_{it}) + \cdots + b_{Mj}(p_t)h_M(m_{it}) \tag{14}$$

where $b_{0j}(p_t), \ldots, b_{Mj}(p_t)$ and $h_1(m_{it}), \ldots, h_M(m_{it})$ are redefined in the obvious way. If we denote individual expenditure weights as $\mu_{it} = m_{it}/E_t(m_{it})$, then aggregate budget shares are

$$\begin{aligned}\frac{E_t(p_{jt}q_{ijt})}{E_t(m_{it})} = E_t(\mu_{it}w_{ijt}) &= b_{0j}(p_t) + b_{1j}(p_t)E_t\big(\mu_{it}h_1(m_{it})\big) + \cdots \\ &\quad + b_{Mj}(p_t)E_t\big(\mu_{it}h_M(m_{it})\big).\end{aligned} \tag{15}$$

The same remarks on recoverability apply here: the individual budget share coefficients $b_{1j}(p_t), \ldots, b_{Mj}(p_t)$ can be identified with aggregate data with sufficient variation in the distributional terms $E_t(\mu_{it}h_1(m_{it})), \ldots, E_t(\mu_{it}h_M(m_{it}))$ over time. As above, aggregation factors can be used to gauge the difference between aggregate shares and individual shares. We have

$$\begin{aligned}\frac{E_t(p_{jt}q_{ijt})}{E_t(m_{it})} = E_t(\mu_{it}w_{ijt}) &= b_{0j}(p_t) + b_{1j}(p_t)\pi_{1t}h_1\big(E_t(m_{it})\big) + \cdots \\ &\quad + b_{Mj}(p_t)\pi_{Mt}h_M\big(E_t(m_{it})\big)\end{aligned} \tag{16}$$

where by construction

$$\pi_{kt} = \frac{E_t(\mu_{it}h_k(m_{it}))}{h_k(E_t(m_{it}))}, \quad k = 1, \ldots, M, \tag{17}$$

are the aggregation factors. These factors give a compact representation of the distributional influences that cause the aggregate model, and the elasticities derived from it, to differ from the individual model.

The budget share form (14) accommodates exact aggregation through the separation of income and price terms in its additive form. As before, when integrability restrictions are applied to (14), the range of possible model specifications is strongly reduced. A particularly strong result is due to Gorman (1981), who showed that homogeneity and symmetry restrictions imply that the rank of the $J \times (M + 1)$ matrix of coefficients $[b_{mj}(p_t)]$ can be no greater than 3. Lau (1977), Lewbel (1991) and others have characterized the full range of possible forms for the income functions.

2.1.3. Aggregation in rank 2 and rank 3 models

Early exact aggregation models were of rank 2 (for a given value of attributes z_{it}). With budget share equations of the form[11]

$$w_{ijt} = b_{0j}(p_t) + b_{1j}(p_t)h_1(m_{it}), \tag{18}$$

[11] This is Muellbauer's (1975) PIGL form.

preferences can be specified that give rise to either the log-form $h_1(m_{it}) = \ln m_{it}$ or the power form $h_1(m_{it}) = m_{it}^{\theta}$. Typically the former is adopted and this produces Engel curves that are the same as those that underlie the Almost Ideal model and the Translog model (without attributes).[12] In this case, aggregate shares have the form

$$\frac{E_t(p_{jt}q_{ijt})}{E_t(m_{it})} = E_t(\mu_{it}w_{ijt}) = b_{0j}(p_t) + b_{1j}(p_t)\pi_{1t}\ln E_t(m_{it}) \tag{19}$$

where the relevant aggregation factor is the following entropy measure for the m_{it} distribution:

$$\pi_{1t} = \frac{E_t(\mu_{it}\ln m_{it})}{\ln E_t(m_{it})} = \frac{E_t(m_{it}\ln m_{it})}{E_t(m_{it})\ln E_t(m_{it})}, \tag{20}$$

where we have recalled that $\mu_{it} = m_{it}/E_t(m_{it})$. The deviation of π_{1t} from unity describes the degree of bias in recovering (individual) price and income elasticities from aggregate data alone.

Distribution restrictions can be used to facilitate computation of the aggregate statistics as well as studying the aggregation factors. For instance, suppose income is log-normally distributed, with $\ln m_{it}$ distributed normally with mean μ_{mt} and variance σ_{mt}^2. The aggregation factor (20) can easily be seen to be

$$\pi_{1t} = 1 + \frac{1}{2(\mu_{mt}/\sigma_{mt}^2) + 1}. \tag{21}$$

To the extent that the log mean and variance are in stable proportion, π_{1t} will be stable. If the log mean is positive, then $\pi_{1t} > 1$, indicating positive bias from using $\ln E_t(m_{it})$.

Distribution restrictions can also facilitate the more modest goal of a stable relationship between aggregate budget shares and aggregate total expenditure. For instance, suppose that the total expenditure distribution obeys

$$E_t(m_{it}\ln m_{it}) = c_1 E_t(m_{it}) + c_2 E_t(m_{it})\ln E_t(m_{it}). \tag{22}$$

Then aggregate budget shares are

$$\frac{E_t(p_{jt}q_{ijt})}{E_t(m_{it})} = b_{0j}(p_t) + b_{1j}(p_t)\big(c_1 + c_2\ln E_t(m_{it})\big) \tag{23}$$

so that a relationship of the form

$$\frac{E_t(p_{jt}q_{ijt})}{E_t(m_{it})} = \tilde{b}_{0j}(p_t) + \tilde{b}_{1j}(p_t)\ln E_t(m_{it}) \tag{24}$$

would describe aggregate data well.

[12] It is worthwhile to note that with the power form, estimation of θ with aggregate data would be complicated, because the aggregation statistics would depend in a complicated way on θ.

Here, integrability properties from individual demands can impart similar restrictions to the aggregate relationship. Lewbel (1991) shows that if individual shares

$$w_{ijt} = b_{0j}(p_t) + b_{1j}(p_t) \ln m_{it} \tag{25}$$

satisfy symmetry, additivity and homogeneity properties, then so will

$$w_{ijt} = b_{0j}(p_t) + b_{1j}(p_t)(\kappa + \ln m_{it}). \tag{26}$$

The analogy of (23) and (26) makes clear that if $c_2 = 1$, then the aggregate model will satisfy symmetry, additivity and homogeneity. As such, some partial integrability restrictions may be applicable at the aggregate level.[13]

As we discuss in Section 2.2 below, rank 2 models of the form (18) fail on empirical grounds. Evidence points to the need for more extensive income effects (for given demographic attributes z_{it}), such as available from rank 3 exact aggregation specifications. In particular, rank 3 budget share systems that include terms in $(\ln m_{it})^2$ (as well as individual attributes) seem to do a good job of fitting the data, such as the QUAIDS system of Banks, Blundell and Lewbel (1997), described further in Section 2.2 below. In these cases, corresponding to the quadratic term $(\ln m_{it})^2$, there will be an additional aggregation factor to examine,

$$\pi_{2t} = \frac{E_t(\mu_{it}(\ln m_{it})^2)}{(\ln E_t(m_{it}))^2} = \frac{E_t(m_{it}(\ln m_{it})^2)}{E_t(m_{it})(\ln E_t(m_{it}))^2}. \tag{27}$$

In analogy to (22), one can define partial distributional restrictions so that aggregate shares are well approximated as a quadratic function of $\ln E_t(m_{it})$.

2.1.4. Heterogeneous attributes

As we noted in our earlier discussion, the empirical analysis of individual-level data has uncovered substantial demographic effects on demand. Here we reintroduce attributes z_{it} into the equations, to capture individual heterogeneity not related to income. Since z_{it} varies across consumers, for exact aggregation, z_{it} must be incorporated in a similar fashion to total expenditure m_{it}. The budget share form (14) is extended generally to

$$w_{ijt} = b_{0j}(p_t) + b_{1j}(p_t)h_1(m_{it}, z_{it}) + \cdots + b_{Mj}(p_t)h_M(m_{it}, z_{it}). \tag{28}$$

Restrictions from integrability theory must apply for each value of the characteristics z_{it}. For instance, Gorman's rank theory implies that the share model can be rewritten

[13] It is tempting to consider the case of $c_1 = 0$, $c_2 = 1$, which would imply that the aggregation factor $\pi_{1t} = 1$ (and no aggregation bias). However, that case appears impossible, although we do not provide a proof. For instance, if m_{it} were lognormally distributed, $c_1 = 0$, $c_2 = 1$ would only occur if $\ln m_{it}$ had zero variance.

with two terms that depend on m_{it}, but there is no immediate limit on the number of h terms that depend only on characteristics z_{it}.[14]

Budget share models that incorporate consumer characteristics in this fashion were first introduced by Jorgenson, Lau and Stoker (1980, 1982). Aggregation factors arise for attribute terms, that necessarily involve interactions between income and attributes. The simplest factors arise for terms that depend only on characteristics, as in $h_j(m_{it}, z_{it}) = z_{it}$, namely

$$\pi_t^z = \frac{E_t(\mu_{it} z_{it})}{E_t(z_{it})} = \frac{E_t(m_{it} z_{it})}{E_t(m_{it}) E_t(z_{it})}. \tag{29}$$

This can be seen as the ratio of the income-weighted mean of z_{it} to the unweighted mean of z_{it}. If z_{it} is an indicator, say $z_{it} = 1$ for households with children and $z_{it} = 0$ for households without children, then π_t^z is the percentage of expenditure accounted for by households with children, $E_t(m_{it} z_{it})/E_t(m_{it})$, divided by the percentage of households with children, $E_t(z_{it})$.

More complicated factors arise with expenditure-characteristic effects; for example, if $h_j(m_{it}, z_{it}) = z_{it} \ln m_{it}$ then the relevant aggregation factor is

$$\pi_{1t}^z = \frac{E_t(\mu_{it} z_{it} \ln m_{it})}{E_t(z_{it}) \ln E_t(m_{it})} = \frac{E_t(m_{it} z_{it} \ln m_{it})}{E_t(m_{it}) E_t(z_{it}) \ln E_t(m_{it})}. \tag{30}$$

As before, in analogy to (22), one can derive partial distributional restrictions so that aggregate shares are well approximated as a function of $E_t(m_{it})$ and $E_t(z_{it})$.

2.2. Empirical evidence and the specification of aggregate demand models

2.2.1. What do individual demands look like?

Demand behavior at the individual household level is nonlinear. As we have mentioned, it is not realistic to assume that demands are linear in total expenditures and relative prices. To illustrate typical shapes of income structure of budget shares, Figures 1 and 2 present estimates of Engel curves of two commodity groups for the demographic group of married couples without children, in the British Family Expenditure Survey (FES).[15] Each figure plots the fitted values of a polynomial (quadratic) regression in log total expenditure, together with a nonparametric kernel regression. We see that for food

[14] A simple linear transformation will not in general be consistent with consumer optimization. Blundell, Browning and Crawford (2003) show that if budget shares have a form that is additive in functions of $\ln m_{it}$ and demographics, then if (i) Slutsky symmetry holds and (ii) the effects of demographics on budget shares are unrestricted then they have to be linear in $\ln m_{it}$.

[15] The FES is a random sample of around 7000 households per year. The commodity groups are nondurable expenditures grouped into: food-in, food-out, electricity, gas, adult clothing, children's clothing and footwear, household services, personal goods and services, leisure goods, entertainment, leisure services, fares, motoring and gasoline. More precise definitions and descriptive statistics are available on request.

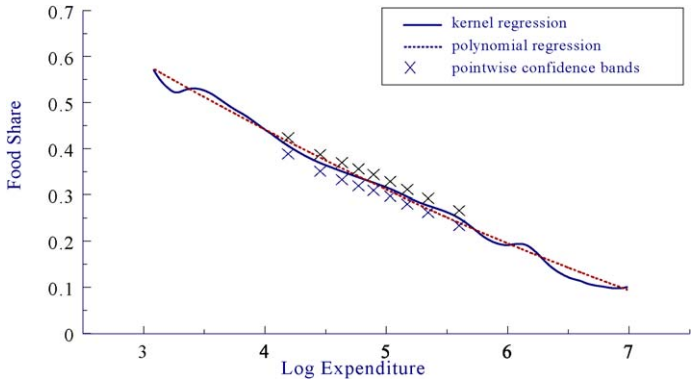

Figure 1. Nonparametric Engel curve: Food share.

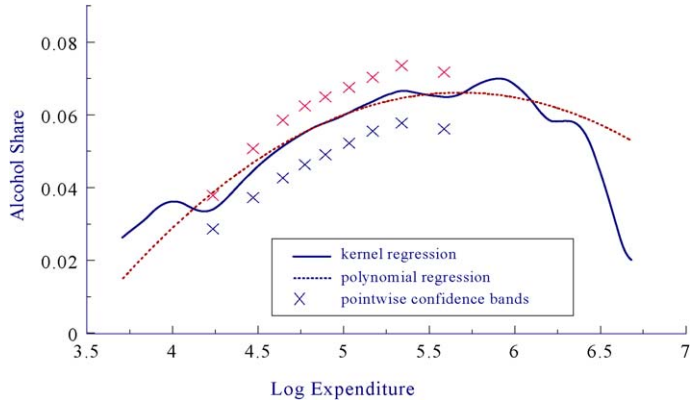

Figure 2. Nonparametric Engel curve: Alcohol share.

expenditures, an equation that expressed the food share as a linear function of log expenditure would be roughly correct. For alcohol expenditures, the income structure is more complex, requiring quadratic terms in log expenditure. Moreover, as one varies the demographic group, the shapes of the analogous Engel curves are similar, but they vary in level and slope.

The QUAIDS model of Banks, Blundell and Lewbel (1997) seems to be sufficiently flexible to capture these empirical patterns. In the QUAIDS model, expenditure shares have the form

$$w_{ijt} = \alpha_j + \gamma_j' \ln p_t + \beta_j \big(\ln m_{it} - \ln a(p_t)\big) + \lambda_j \frac{(\ln m_{it} - \ln a(p_t))^2}{c(p_t)} + u_{ijt} \tag{31}$$

where $a(p_t)$ and $c(p_t)$ are given as

$$\ln a(p_t) = \alpha' \ln p_t + \frac{1}{2} \ln p_t' \Gamma \ln p_t,$$

$$\ln c(p_t) = \beta' \ln p_t,$$

with $\alpha = (\alpha_1, \ldots, \alpha_N)'$, $\beta = (\beta_1, \ldots, \beta_N)'$, $\lambda = (\lambda_1, \ldots, \lambda_N)'$ and

$$\Gamma = \begin{pmatrix} \gamma_1' \\ \vdots \\ \gamma_N' \end{pmatrix}.$$

This generalizes the (linear) Almost Ideal demand system by allowing nonzero λ_i values, with the denominator $c(p_t)$ required to maintain the integrability restrictions. Banks, Blundell and Lewbel (1997) do extensive empirical analysis and establish the importance of the quadratic log expenditure terms for many commodities. Interestingly, they find no evidence of the rejection of integrability restrictions associated with homogeneity or symmetry.

To include demographic attributes, an attractive specification is the 'shape-invariant' specification of Blundell, Duncan and Pendakur (1998). Suppose that $g_j^0(\ln m_i)$ denotes a 'base' share equation, then a shape-invariant model specifies budget shares as

$$w_{ijt} = g_j^0\big(\ln m_i - \phi\big(z_{it}'\theta\big)\big) + z_{it}'\varphi_j.$$

The shape-invariant version of the QUAIDS model allows demographic variation in the α_j terms. In Banks, Blundell and Lewbel (1997), the α_j, β_j and λ_j terms in (31) are allowed to vary with many attributes z_{it}.[16] Family size, family composition, labor market status, occupation and education are all found to be important attributes for many commodities.[17]

2.2.2. *The implications for aggregate behavior*

The stability and interpretation of aggregate relationships can be assessed from examining the appropriate aggregation factors. We can compute the empirical counterparts to the factors by replacing expectations with sample averages. For instance, π_{1t} of (20) is estimated as

$$\hat{\pi}_{1t} = \frac{\sum_i (\hat{\mu}_{it} \ln m_{it})/n_t}{\ln(\sum_i m_{it}/n_t)} \tag{32}$$

[16] For instance, $\alpha_j + \delta_j' z_{it}$ is used in place of α_j, and similar specifications for β_j and λ_j terms.

[17] Various methods can be used to estimate the QUAIDS model, with the iterated moment estimator of Blundell and Robin (2000) particularly straightforward. Banks, Blundell and Lewbel (1997) deal with endogeneity of total expenditures, using various instruments. Finally, we note that Jorgenson and Slesnick (2005) have recently combined a Translog demand model (of rank 3) with an intertemporal allocation model, to model aggregate demand and labor supply in the United States.

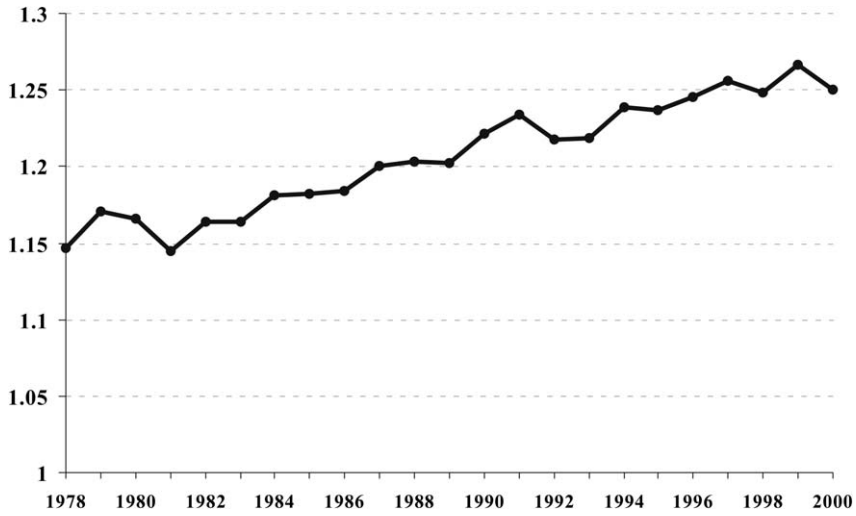

Figure 3. Aggregation factor for households with children.

and π_t^z of (29) is estimated as

$$\hat{\pi}_t^z = \frac{\sum_i \hat{\mu}_{it} z_{it}/n_t}{\sum_i z_{it}/n_t} \tag{33}$$

where we recall that the weights have the form $\hat{\mu}_{it} = m_{it}/(\sum_i m_{it}/n_t)$. Similarly, quadratic terms in $\ln m_{it}$ will require the analysis of the empirical counterpart to the term (27). Interactions of the β and λ terms in (31) with demographic attributes necessitate examination of the empirical counterparts of terms of the form (30). We can also study aggregation factors computed over different subgroups of the population, to see how aggregate demand would vary over those subgroups.

Figure 3 presents the estimated π_t^z term for the impact of children on household demands. This shows a systematic rise in the share of nondurable expenditures attributable to families with children over the 1980s and 1990s. The aggregate bias associated with using observed percentage of households with children (as opposed to the income distribution across households with and without children) varies from 15% to 25%. The path of π_t^z also follows the UK business cycle and the path of aggregate expenditure with downturns in 1981 and 1992.

Figure 4 presents the estimated π_{1t} and π_{2t} terms relating to the $\ln m_{it}$ and $(\ln m_{it})^2$ expressions in the QUAIDS demand model. It is immediately clear that these also display systematic time-series variation, but in comparison to π_t^z above, they increase over the first period of our sample and fall towards the end. The bias in aggregation exhibited for the $(\ln m_{it})^2$ term is more than double that exhibited for the $\ln m_{it}$ term.

Figure 5 presents the aggregation factors for the $\ln m_{it}$ term delineated by certain household types. The baseline $\ln m$ line is the same as that in Figure 4. The other two

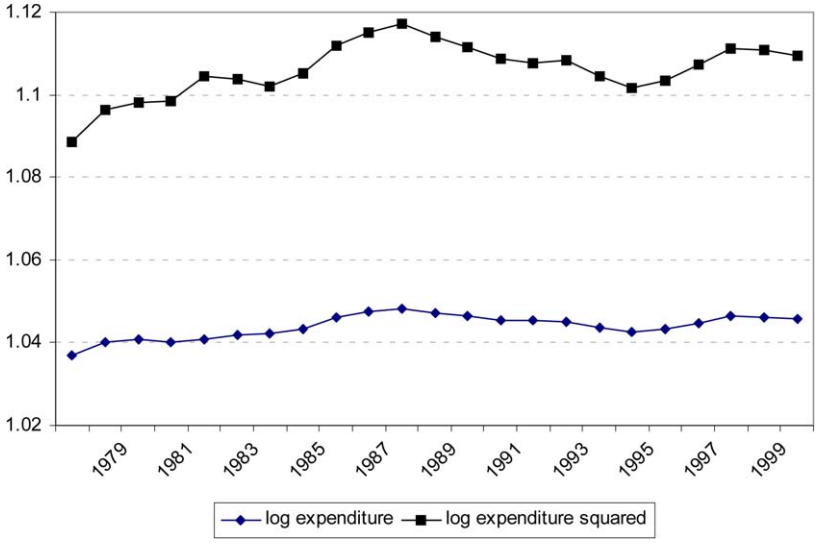

Figure 4. Aggregation factors for income structure.

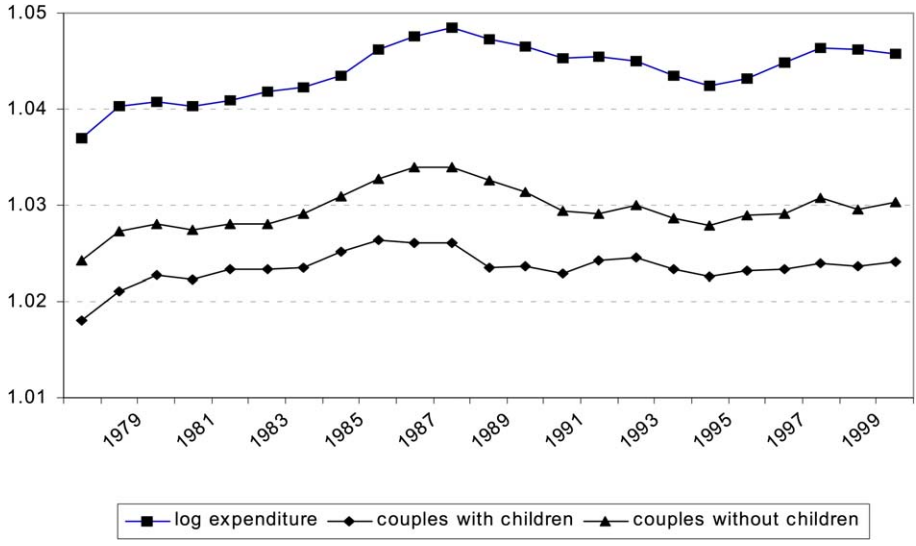

Figure 5. Aggregation factors for income structure by certain household types.

lines correspond to interactions for couples without children and for couples with children. While the time patterns of aggregation factors are similar, they are at different

levels, indicating different levels of bias associated with aggregation over these subgroups.

Finally, it is worthwhile to mention some calculations we carried out on whether distributional restrictions such as (22) are capable of representing the aggregate movements in total expenditure data. Using the time-series of distributional statistics from the FES data, we followed Lewbel (1991) and implemented each of these approximations as a regression. With demographic interaction terms, the aggregate model will only simplify if these conditions also apply to each demographic subgroup. In virtually every case, we found the fit of the appropriate regressions to be quite close (say R^2 in the region of 0.99). This gives support to the idea that aggregate demand relationships are reasonably stable empirically. However, the evidence on the c_j terms implies that aggregation factors are substantially different from one, so again, estimates of the price and income elasticities using aggregate data alone will be not be accurate.

2.3. Aggregation of demand without individual structure

We close this section with discussion of a nontraditional approach given in Hildenbrand (1994), which is to study specific aspects of aggregate demand structure without relying on assumptions on the behavior of individual consumers. This work makes heavy use of empirical regularities in the observed distribution of consumer expenditures across the population.

We can understand the nature of this approach from a simple example. Suppose we are interested in whether aggregate demand for good j decreases when price p_j increases (obeying the "Law of Demand"), and we omit reference to other goods and time t for simplicity. Denote the conditional expectation of demand q_{ij} for good j, given income m_i and price p_j, as $g_j(p_j, m_i)$. Aggregate demand for good j is

$$E(q_{ij}) = G(p_j) = \int g(p_j, m)\, \mathrm{d}F(m)$$

and our interest is in whether $\mathrm{d}G/\mathrm{d}p_j < 0$. Form this derivative, applying the Slutsky decomposition to $g(p_j, m)$ as

$$\begin{aligned}\frac{\mathrm{d}G}{\mathrm{d}p_j} &= \int \left[\left.\frac{\mathrm{d}g}{\mathrm{d}p_j}\right|_{\mathrm{comp}} - g(p_j, m)\frac{\mathrm{d}g}{\mathrm{d}m} \right] \mathrm{d}F(m) \\ &= \int \left.\frac{\mathrm{d}g}{\mathrm{d}p_j}\right|_{\mathrm{comp}} \mathrm{d}F(m) - \int g(p_j, m)\frac{\mathrm{d}g}{\mathrm{d}m}\, \mathrm{d}F(m) \\ &= S - A. \end{aligned} \tag{34}$$

The price effect on aggregate demand decomposes into the mean compensated price effect S and the mean income effect A. If we take S as negative, which is fairly uncontroversial, then we know that $\mathrm{d}G/\mathrm{d}p_j < 0$ if the income effect $A > 0$. Looking once more at A, we can see various ways of ascertaining whether $A > 0$:

$$A = \int g(p_j, m)\frac{\mathrm{d}g}{\mathrm{d}m}\, \mathrm{d}F(m) = \frac{1}{2}\int \frac{\mathrm{d}[g(p_j, m)^2]}{\mathrm{d}m}\, \mathrm{d}F(m). \tag{35}$$

Without making any structural assumptions on $g(p_j, m)$, one could estimate A with the first expression using nonparametric estimates of $g(\cdot)$ and its derivative. Or, we could examine directly whether the "spread" $g(p_j, m)^2$ is increasing with m, and if so, conclude that $A > 0$.

This gives the flavor of this work without doing justice to the details. The main contribution is to link up properties of aggregate demand directly with aspects of the distribution of demands across the population. Hildenbrand (1998) shows that increasing spread is a common phenomenon in data on British households, and it is likely to be valid generally. More broadly, this work has stimulated extensive study of the distribution of household expenditures, with a different perspective from traditional demand modeling. Using nonparametric methods, Härdle, Hildenbrand and Jerison (1991) study aggregate income effects across a wide range of goods, and conclude that the "law of demand" likely holds quite generally. Hildenbrand and Kneip (1993) obtain similar findings on income structure by directly examining the dimensionality of vectors of individual demands.[18] See Hildenbrand (1994) for an overview of this work, as well as Hildenbrand (1998) for an examination of variations in the British expenditure distribution within a similar framework.

3. Consumption and wealth

We now turn to a discussion of total consumption expenditures. Here the empirical problem is to characterize consumption expenditures over time periods, including how they relate to income and wealth. The individual level is typically that of a household (or an individual person, depending on data source). The economic aggregate to be modeled is average consumption expenditures over time, and we are interested in how aggregate consumption and saving relate to income and wealth across the economy, as well as to interest rates. This relationship is essential for understanding how interest rates will evolve as the population changes demographically, for instance.

Consumption expenditures are determined through a forward-looking plan, that takes into account the needs of individuals over time, as well as uncertainty in wealth levels. There is substantial evidence of demographic effects and nonlinearities in consumption at the individual level,[19] so we will need to consider heterogeneity in tastes as before. Accordingly, aggregate consumption is affected by the structure of households and especially the age distribution, and will also be affected by inequality in the distribution of wealth. We are not concerned here with heterogeneity in market participation *per se*, as everyone has nonzero consumption expenditures. Later we discuss some issues

[18] This is related to the transformation modeling structure of Grandmont (1992). It is clear that the dimensionality of exact aggregation demand systems is given by the number of independent income/attribute terms [cf. Diewert (1977) and Stoker (1984b)].

[19] See Attanasio and Weber (1993a, 1993b) and Attanasio and Browning (1995), among many others.

raised by liquidity constraints, which have much in common with market participation modeling as described in Section 4.

Our primary focus is on heterogeneity with regard to risks in income and wealth levels, and how the forward-planning process is affected by them. We take into account the nature of the income and wealth shocks, as well as the nature of the credit markets that provide insurance against negative shocks.

We consider four different types of shocks, delineated by whether the effects are permanent or transitory, and whether they are aggregate, affecting all consumers, or individual in nature. Aggregate permanent shocks can refer to permanent changes in the productive capability of the economy – such as running out of a key natural resource, or skill-biased technical change – as well as to permanent changes in taxes or other policies that affect saving. Individual permanent shocks include permanent changes in an individual's ability to earn income, such as chronic bad health and long-term changes in type and status of employment. Aggregate transitory shocks refer to temporary aggregate phenomena, such as exchange rate variation, bad weather and so forth. Individual transitory shocks include temporary job lay-offs, temporary illnesses, etc. Many different situations of uncertainty can be accounted for by combinations of these four different types of shocks.

In terms of risk exposure and markets, there are various scenarios to consider. With complete markets, all risks are insured, and an individual's consumption path is unaffected by the evolution of the individual's income over time.[20] When markets are not complete, the extent of available insurance markets becomes important, and determines the degree to which different individual risks are important for aggregate consumption behavior. For example, in the absence of credit market constraints, idiosyncratic risks may be open to self-insurance. But in that case there may be little insurance available for aggregate shocks or even for permanent idiosyncratic shocks. Our discussion takes into account the type of income risks and how risk exposure affects aggregate consumption.

Most of our discussion focuses on individual consumption plans and their implications for aggregate consumption. Beyond this, we can consider the feedback effects on consumption and wealth generated through general equilibrium. For instance, if a certain group of consumers systematically saves more than others, then in equilibrium those consumers will be wealthier, and their saving behavior will be a dominant influence on the evolution of aggregate wealth. The study of this important topic is in its infancy, and has been analyzed primarily with calibrated macroeconomic growth models. We include a discussion of some of this work.

3.1. Consumption growth and income shocks

In our framework, in each period t, individual i chooses consumption expenditures c_{it} by maximizing expected utility subject to an asset accumulation constraint. Individual

[20] See Atkeson and Ogaki (1996) for a model of aggregate expenditure allocation over time and to individual goods based on addilog preferences, assuming that complete markets exist.

i has heterogeneous attributes z_{it} that affect preferences. There is a common, riskless interest rate r_t. We assume separability between consumption and labor supply in each time period, and separability of preferences over time.

We begin with a discussion of aggregation with quadratic preferences. This allows us to focus on the issues of different types of income shocks and insurance, without dealing with nonlinearity. In Section 3.2, we consider more realistic preferences that allow precautionary saving.

When individual within-period utilities are quadratic in current consumption, we have the familiar certainty-equivalent formulation in which there is no precautionary saving. Within-period utilities are given as

$$U_{it}(c_{it}) = -\frac{1}{2}(a_{it} - c_{it})^2 \tag{36}$$

for $c_{it} < a_{it}$. We model individual heterogeneity by connecting a_{it} to individual attributes as

$$a_{it} = \alpha + \beta' z_{it}. \tag{37}$$

With the discount rate equal to the real interest rate, maximizing the expected sum of discounted utilities gives the following optimal plan for the consumer [Hall (1978)]:

$$\Delta c_{it} = \Delta\alpha_{it} + \xi_{it} = \beta' \Delta z_{it} + \xi_{it}. \tag{38}$$

Defining $\Omega_{i,t-1}$ as the information set for individual i in period $t-1$, the consumption innovation ξ_{it} obeys

$$E[\xi_{it}|\Omega_{i,t-1}] = 0. \tag{39}$$

In what follows we will use a time superscript to denote this conditional expectation, namely $E^{t-1}(\cdot) \equiv E[\cdot|\Omega_{i,t-1}]$ to distinguish it from the population average in period t (which uses a time subscript as in $E_t(\cdot)$). Notice, the model (38) is linear in the change in attributes Δz_{it} with constant coefficients β, plus the consumption innovation. In other words, this model is in exact aggregation form with regard to the attributes z_{it} that affect preferences.

3.1.1. Idiosyncratic income variation and aggregate shocks

When the only uncertainty arises from real income, the consumption innovation ξ_{it} can be directly related to the stochastic process for income. We begin by spelling out the income process in a meaningful way. Express income y_{it} as the sum of transitory and permanent components

$$y_{it} = y_{it}^P + y_{it}^T \tag{40}$$

and assume that the transitory component is serially independent. We assume that the permanent component follows a random walk

$$y_{it}^P = y_{it-1}^P + \eta_{it}^P, \tag{41}$$

where the innovation η_{it}^P is serially independent.

Next, decompose these two components into a common aggregate effect and an idiosyncratic effect

$$\eta_{it}^P = \eta_t + \varepsilon_{it}, \tag{42}$$

$$y_{it}^T = u_t + v_{it}. \tag{43}$$

Here η_t is the common aggregate permanent shock, ε_{it} is the permanent shock at the individual level, u_t is the aggregate transitory shock and v_{it} is the individual transitory shock – the four types of income shocks discussed above. This mixture of permanent and transitory shocks has been found to provide a good approximation to the panel data process for log incomes; see MaCurdy (1982) and Meghir and Pistaferri (2004). We assume that the individual shocks are normalized to average to zero across the population, namely $E_t(\varepsilon_{it}) = 0$ and $E_t(v_{it}) = 0$.

The stochastic process for individual income then takes the form

$$\Delta y_{it} = \eta_t + \varepsilon_{it} + \Delta u_t + \Delta v_{it}. \tag{44}$$

The stochastic process for aggregate income has the form

$$\Delta E_t(y_{it}) = \eta_t + \Delta u_t \tag{45}$$

where, again, E_t denotes expectation (associated with averaging) across the population of agents at time t.

3.1.2. Income shocks and insurance

The first scenario is where individual (and aggregate) shocks are not insurable. Here the optimal consumption innovation ξ_{it} for the individual will adjust fully to permanent income shocks but only adjust to the annuity value of transitory shocks. To see this, again suppose that real interest rates are constant and equal the discount rate. Under quadratic preferences (36), consumption growth can be written [Deaton and Paxson (1994)] as

$$\Delta c_{it} = \beta' \Delta z_{it} + \eta_t + \varepsilon_{it} + \tau_t(\Delta u_t + \Delta v_{it}), \tag{46}$$

where τ_t is the annuitization rate for a transitory shock with planning over a finite horizon.[21] Clearly, expected growth is determined by preference attributes as

$$E^{t-1}(\Delta c_{it}) = E(\Delta c_{it} | \Omega_{i,t-1}) = \beta' \Delta z_{it}. \tag{47}$$

[21] If L is the time horizon, then $\tau_t = r/[(1+r)(1-(1+r)^{-(L-t+1)})]$. Clearly $\tau_t \to 0$ as $r \to 0$. Note that for a small interest rate, we have $\tau_t \approx 0$, so that the transitory shocks become irrelevant for consumption growth.

Aggregate consumption has the form

$$\Delta E_t(c_{it}) = \beta' \Delta E_t(z_{it}) + \eta_t + \tau_t \Delta u_t. \tag{48}$$

Thus, the aggregate data are described exactly by a representative agent model with quadratic preferences and characteristics $E_t(z_{it})$ facing a permanent/transitory income process.[22]

For the second scenario, suppose individual shocks can be fully insured, either through informal processes or through credit markets. Now individual consumption growth depends only on aggregate shocks

$$\Delta c_{it} = \beta' \Delta z_{it} + \eta_t + \tau_t \Delta u_t. \tag{49}$$

Consequently, with (48), we will have

$$\Delta c_{it} = \beta' \big(\Delta z_{it} - \Delta E_t(z_{it})\big) + \Delta E_t(c_{it}). \tag{50}$$

Thus, consumption growth at the individual level equals aggregate consumption growth plus an adjustment for individual preferences.

Finally, the third scenario is where all shocks (aggregate and individual) are fully insurable. Now individual consumption growth will be the planned changes $\beta' \Delta z_{it}$ only, and aggregate consumption growth will be the mean of those changes $\beta' \Delta E_t(z_{it})$. This is the most complete "representative agent" case, as complete insurance has removed the relevance of all income risks.

3.1.3. Incomplete information

It is interesting to note that in our simplest framework, incomplete information can cause aggregate consumption to fail to have random walk structure. In particular, suppose individual shocks are not completely insurable and consumers cannot distinguish between individual and aggregate shocks. To keep it simple, also assume that there are no varying preference attributes z_{it}. Following Pischke (1995), individual i will view the income process (44) as an MA(1) process:

$$\Delta y_{it} = \zeta_{it} - \theta \zeta_{it-1}, \tag{51}$$

where the θ parameter is a function of the relative variances of the shocks.

Changes in individual consumption are simply

$$\Delta c_{it} = (1 - \theta)\zeta_{it}.$$

[22] Aside from the drift term $\beta' \Delta E_t(z_{it})$, aggregate consumption is a random walk. In particular, the orthogonality conditions $E^{t-1}(\eta_t + \tau_t \Delta u_t) = E(\eta_t + \tau_t \Delta u_t | \Omega_{i,t-1}) = 0$ hold at the individual level and therefore also hold at the aggregate level.

Note that it is still the case that $E^{t-1}(\Delta c_{it}) = E(\Delta c_{it}|\Omega_{i,t-1}) = 0$. However, from (51) we have that

$$\Delta c_{it} - \theta \Delta c_{it-1} = (1-\theta)\Delta y_{it}. \tag{52}$$

Replacing Δy_{it} by (44) and averaging over consumers we find

$$\Delta E_t(c_{it}) = \theta \Delta E_{t-1}(c_{it-1}) + (1-\theta)(\eta_t + \Delta u_t) \tag{53}$$

so that aggregate consumption is clearly not a random walk.

3.2. Aggregate consumption growth with precautionary saving

With quadratic preferences, consumption growth can be written as linear in individual attributes – in exact aggregation form – and we are able to isolate the impacts of different kinds of income shocks and insurance scenarios. To allow for precautionary saving, we must also account for nonlinearity in the basic consumption process. For this, we now consider the most standard consumption model used in empirical work, that based on Constant Relative Risk Aversion (CRRA) preferences.

3.2.1. Consumption growth with CRRA preferences

We assume that within-period utility is

$$U_{it}(c_{it}) = e^{a_{it}} \left[\frac{c_{it}^{1-\frac{1}{s_{it}}}}{1-\frac{1}{s_{it}}} \right], \tag{54}$$

where a_{it} permits scaling in marginal utility levels (or individual subjective discount rates), and s_{it} is the intertemporal elasticity of substitution, reflecting the willingness of individual i to trade off today's consumption for future consumption. As before, we will model the heterogeneity in a_{it} and s_{it} via individual attributes z_{it}.

We now adopt a multiplicative stochastic income process, with the decomposition expressed in log form as

$$\Delta \ln y_{it} = \eta_t + \varepsilon_{it} + \Delta u_t + \Delta v_{it}. \tag{55}$$

The permanent and transitory error components in the income process are decomposed into aggregate and individual terms, as in (44). As noted before, this income growth specification is closely in accord with the typical panel data models of income or earnings, and it will neatly complement our equations for consumption growth with CRRA preferences. In addition, we assume that the interest rate r_t is small, for simplicity, and is not subject to unanticipated shocks.

With precautionary saving, consumption growth depends on the conditional variances of the uninsurable components of shocks to income. Specifically, with CRRA preferences (54) and log income process (55), we have the following log-linear approximation

for consumption growth[23]

$$\Delta \ln c_{it} = \rho r_t + (\beta + \varphi r_t)' z_{it} + k_1 \sigma_{it}^{t-1} + k_2 \sigma_{At}^{t-1} + \kappa_1 \varepsilon_{it} + \kappa_2 \eta_t, \tag{56}$$

where σ_{it}^{t-1} is the conditional variance of idiosyncratic risk (conditional on $t-1$ information $\Omega_{i,t-1}$) and σ_{At}^{t-1} is the conditional variance of aggregate risk. The attributes z_{it} represent the impact of heterogeneity in a_{it}, or individual subjective discount rates, and the intertemporal elasticity of substitution $s_{it} = \rho + \beta + \varphi' z_{it}$. Typically in empirical applications, z_{it} will include levels and changes in observable attributes, and unobserved factors may also be appropriate.[24] As before,

$$E(\varepsilon_{it} | \Omega_{i,t-1}) = E^{t-1}(\varepsilon_{it}) = 0, \tag{57}$$

$$E(\eta_t | \Omega_{i,t-1}) = E^{t-1}(\eta_t) = 0. \tag{58}$$

To sum up, in contrast to the quadratic preference case, the growth equation (56) is nonlinear in consumption, and it includes conditional variance terms which capture the importance of precautionary saving.

A consistent aggregate of the individual model (56) is given by

$$E_t(\Delta \ln c_{it}) = \rho r_t + (\beta + \varphi r_t)' E_t(z_{it}) + k_1 E_t(\sigma_{it}^{t-1}) + k_2 \sigma_{At}^{t-1} + \kappa_2 \eta_t, \tag{59}$$

where $E_t(\Delta \ln c_{it})$ refers to the population mean of the cross-section distribution of $\Delta \ln c_{it}$ in period t, and so on. The t subscript again refers to averaging across the population of consumers, and we have normalized $E_t(\varepsilon_{it}) = 0$ as before. Provided $E_t(\ln c_{it-1}) = E_{t-1}(\ln c_{it-1})$, Equation (59) gives a model of changes over time in $E_t(\ln c_{it})$, which is a natural aggregate given the log form of the model (56).

However, $E_t(\ln c_{it})$ is not the aggregate typically observed nor is it of much policy interest. Of central interest is per-capita consumption $E_t(c_{it})$ or total consumption $n_t E_t(c_{it})$. Deriving an equation for the appropriate aggregates involves dealing with the 'log' nonlinearity, to which we now turn.[25]

3.2.2. How is consumption distributed?

Since the individual consumption growth equations are nonlinear, we must make distributional assumptions to be able to formulate an equation for aggregate consumption.

[23] See Blundell and Stoker (1999) for a precise derivation and discussion of this approximation.

[24] See Banks, Blundell and Brugiavini (2001) for a detailed empirical specification of consumption growth in this form.

[25] If we evaluate the individual model at aggregate values, we get

$$\Delta \ln E_t(c_{it}) = \rho r_t + (\beta + \varphi r_t)' E_t(z_{it}) + k_2 \sigma_{At}^{t-1} + \omega_t.$$

Here ω_t is a 'catch-all' term containing the features that induce aggregation bias, that will not satisfy the orthogonality condition $E^{t-1}(\omega_t) = 0$. It is also worthwhile to note that empirical models of aggregate consumption also typically omit the terms $E_t(z_{it})$ and $k_2 \sigma_{At}^{t-1}$.

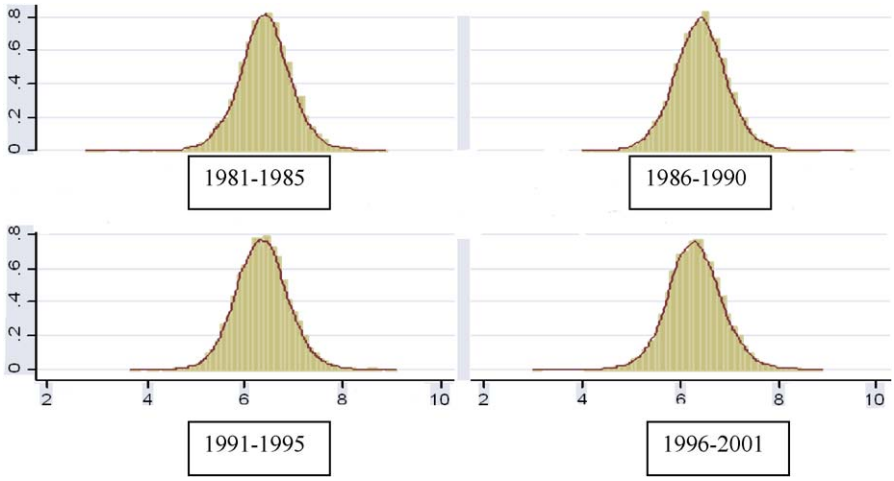

Figure 6. The distribution of log nondurable consumption expenditure: US 1981–2001.

In the following, we will assume lognormality of various elements of the consumption process. Here we point out that this is motivated by an important empirical regularity – namely, individual consumption does appear to be lognormally distributed, at least in developed countries such as the United States and the United Kingdom.

Figure 6 shows the distribution of log-consumption using US consumer expenditure data across the last two decades. Consumption is taken as real expenditure on nondurables and services, and is plotted by five-year bands to achieve a reasonable sample size. Each log-consumption distribution has a striking resemblance to a normal density. In the experience of the authors, this result is often replicated in more disaggregated data by year and various demographic categorizations, such as birth cohort, and also in other countries including in the Family Expenditure Survey data for the UK. Given this regularity, one would certainly start with lognormality assumptions such as those we make below, and any subsequent refinements would need to preserve normality of the marginal distribution of log-consumption.

3.2.3. Insurance and aggregation with precautionary saving

As with our previous discussion, we must consider aggregation under different scenarios of insurance for income risks. We again assume that agents have the same information set, namely $\Omega_{i,t-1} = \Omega_{t-1}$ for all i, t.

We begin with the scenario in which there is full insurance for individual risks, or pooling of idiosyncratic risk across individuals. Here insurance and credit markets are sufficiently complete to remove individual risk terms in individual income and consumption streams, so $\varepsilon_{it} = 0$ and $\sigma_{it}^{t-1} = 0$ for all i, t. The individual model (56)

becomes

$$\Delta \ln c_{it} = \rho r_t + (\beta + \varphi r_t)' z_{it} + k_2 \sigma_{At}^{t-1} + \kappa_2 \eta_t \tag{60}$$

with $E^{t-1}(\eta_t) = 0$. The mean-log model (59) is now written as

$$E_t(\ln c_{it}) - E_t(\ln c_{it-1}) = \rho r_t + (\beta + \varphi r_t)' E_t(z_{it}) + k_2 \sigma_{At}^{t-1} + \kappa_2 \eta_t. \tag{61}$$

The relevant aggregate is per-capita consumption $E_t(c_{it})$. Per-capita consumption is given by

$$\begin{aligned} E_t(c_{it}) &= E_t\big[\exp\big(\ln c_{it-1} + \rho r_t + (\beta + \varphi r_t)' z_{it} + k_2 \sigma_{At}^{t-1} + \kappa_2 \eta_t\big)\big] \\ &= \exp\big(\rho r_t + k_2 \sigma_{At}^{t-1} + \kappa_2 \eta_t\big) \cdot E_t\big[c_{it-1} \exp\big((\beta + \varphi r_t)' z_{it}\big)\big] \end{aligned} \tag{62}$$

with the impact of log-linearity arising in the final term, a weighted average of attribute terms interacted with lagged consumption c_{it-1}.

Of primary interest is aggregate consumption growth, or the log-first-difference in aggregate consumption

$$\Delta \ln E_t(c_{it}) = \ln\left(\frac{E_t(c_{it})}{E_{t-1}(c_{it-1})}\right).$$

This is expressed as

$$\begin{aligned} \Delta \ln E_t(c_{it}) &= \rho r_t + k_2 \sigma_{At}^{t-1} + \kappa_2 \eta_t \\ &\quad + \ln\left(\frac{E_t[c_{it-1} \exp((\beta + \varphi r_t)' z_{it})]}{E_t(c_{it-1})}\right) + \ln\left(\frac{E_t(c_{it-1})}{E_{t-1}(c_{it-1})}\right). \end{aligned} \tag{63}$$

Aggregate consumption growth reflects the interest and risk terms that are common to all consumers, a weighted average of attribute terms, and the log-difference in the average of c_{it-1} at time t versus time $t-1$.

Notice first that even if z_{it} is normally distributed, we cannot conclude that $\ln c_{it}$ is normal. We also need (as a sufficient condition) that $\ln c_{it-1}$ is normal at time t to make such a claim. This would further seem to require normality of $\ln c_{it-2}$ at $t-1$, and so forth into the distant past. In any case, we cover this situation with the broad assumption:

The distribution of c_{it-1} is the same in periods $t-1$ and t. (64)

That is, the population could grow or shrink, but the distribution of c_{it-1} is unchanged. Under that assumption, we can drop the last term in (63)

$$\ln\left(\frac{E_t(c_{it-1})}{E_{t-1}(c_{it-1})}\right) = 0. \tag{65}$$

Lagging the individual model (60) gives an equation for c_{it-1}, but there is no natural way to incorporate that structure directly into the equation for aggregate current

consumption $E_t(c_{it})$.[26] Therefore, we further assume

$$\begin{pmatrix} \ln c_{it-1} \\ \theta_t' z_{it} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{c_{-1},t} \\ \theta_t' E_t(z_{it}) \end{pmatrix}, \begin{bmatrix} \sigma^2_{c_{-1},t} & \Sigma'_{zc_{-1},t}\theta_t \\ \theta_t' \Sigma_{zc_{-1},t} & \theta_t' \Sigma_{zz,t}\theta_t \end{bmatrix} \right), \tag{66}$$

where we have set $\theta_t = (\beta + \varphi r_t)$. This assumption says that

$$\ln c_{it-1} + \theta_t' z_{it} \sim \mathcal{N}\big(\mu_{c_{-1},t} + \theta_t' E_t(z_{it}), \sigma^2_{c_{-1},t} + \theta_t' \Sigma_{zz,t}\theta_t + 2\theta_t' \Sigma_{zc_{-1},t}\big) \tag{67}$$

and

$$\ln c_{it-1} \sim \mathcal{N}\big(\mu_{c_{-1},t}, \sigma^2_{c_{-1},t}\big). \tag{68}$$

We can now solve for an explicit solution to (63): apply (65), (67) and (68) and rearrange to get

$$\begin{aligned} \Delta \ln E_t(c_{it}) = {} & \rho r_t + (\beta + \varphi r_t)' E_t(z_{it}) + k_2 \sigma^{t-1}_{At} + \kappa_2 \eta_t \\ & + \frac{1}{2}\big[(\beta + \varphi r_t)' \Sigma_{zz,t} (\beta + \varphi r_t) + 2(\beta + \varphi r_t)' \Sigma_{zc_{-1},t}\big]. \end{aligned} \tag{69}$$

This is the aggregate model of interest, expressing growth in per-capita consumption as a function of the mean of z, the conditional variance terms from income risk, and the covariances between attributes z and lagged consumption c_{it-1}. This shows how individual heterogeneity manifests itself in aggregate consumption through distributional variance terms. These variance terms vary with r_t if the intertemporal elasticity of substitution varies over the population.

Now consider the scenario where some individual risks are uninsurable. This reintroduces terms ε_{it} and σ^{t-1}_{it} in consumption growth at the individual level, and we must be concerned with how those permanent risks are distributed across the population. In particular, we assume in each period that each individual draws idiosyncratic risk from a common conditional distribution, so that $\sigma^{t-1}_{it} = \sigma^{t-1}_{It}$ for all i. The individual consumption growth equation (56) now appears as

$$\Delta \ln c_{it} = \rho r_t + (\beta + \varphi r_t)' z_{it} + k_1 \sigma^{t-1}_{It} + k_2 \sigma^{t-1}_{At} + \kappa_1 \varepsilon_{it} + \kappa_2 \eta_t. \tag{70}$$

The mechanics for aggregation within this formulation are similar to the previous case, including the normalization $E_t(\varepsilon_{it}) = 0$, but we need to deal explicitly with how the permanent individual shocks ε_{it} covary with $\ln c_{it-1}$. As above, we adopt a stability assumption (64). We then extend (66) to assume that $(\ln c_{it-1}, (\beta + \varphi r_t)' z_{it}, \varepsilon_{it})$ is joint normally distributed. The growth in aggregate average consumption is now given by

$$\Delta \ln E_t(c_{it}) = \rho r_t + (\beta + \varphi r_t)' E_t(z_{it}) + k_1 \sigma^{t-1}_{It} + k_2 \sigma^{t-1}_{At} + \kappa_2 \eta_t + \frac{1}{2}(\Lambda_t) \tag{71}$$

[26] This is because of the potential dependence of c_{it-1} on the same factors as c_{it-2}, and so forth.

where

$$\Lambda_t = (\beta + \varphi r_t)' \Sigma_{zz,t} (\beta + \varphi r_t) + \kappa_1^2 \sigma_{\varepsilon,t}^2 + 2(\beta + \varphi r_t)' \Sigma_{zc_{-1},t} + 2\kappa_1 \sigma_{\varepsilon c_{-1},t} + 2\kappa_1 \Sigma_{\varepsilon z,t} (\beta + \varphi r_t).$$

While complex, this formulation underlines the importance of the distribution of risk across the population. In contrast to the full information model (69), there is a term $\sigma_{\varepsilon,t}^2$ in Λ_t that reflects the changing variance in consumption growth. The term σ_{It}^{t-1} captures how idiosyncratic risk varies, based on $t-1$ information.

We have not explicitly considered unanticipated shocks to the interest rate r_t, or heterogeneity in rates across individuals.[27] Unanticipated shocks in interest would manifest as a correlation between r_t and aggregate income shocks, and would need treatment via instruments in estimation. Heterogeneity in rates could, in principle, be accommodated as with heterogeneous attributes. This would be especially complicated if the overall distributional structure were to shift as interest rates increased or decreased.

3.3. Empirical evidence on aggregating the consumption growth relationship

There are two related aspects of empirical research that are relevant for our analysis of aggregation in consumption growth models. The first concerns the evidence on full insurance of individual risks. How good an approximation would such an assumption be? To settle this, we need to examine whether there is evidence of risk pooling across different individuals and different groups in the economy. For example, does an unexpected change in pension rights, specific to one cohort or generation, get smoothed by transfers across generations? Are idiosyncratic health risks to income fully insured? Even though we may be able to cite individual cases where this perfect insurance paradigm clearly fails, is it nonetheless a reasonable approximation when studying the time-series of aggregate consumption?

The second aspect of empirical evidence concerns the factors in the aggregate model (71) that are typically omitted in studies of aggregate consumption. From the point of view of estimating the intertemporal elasticity parameter ρ, how important are these aggregation factors? How well do they correlate with typically chosen instruments and how likely are they to contaminate tests of excess sensitivity performed with aggregate data?

3.3.1. Evidence on full insurance and risk pooling across consumers

If the full insurance paradigm is a good approximation to reality, then aggregation is considerably simplified and aggregate relationships satisfying the standard optimality conditions can be derived with various conditions on individual preferences. There is a reasonably large and expanding empirical literature on the validity of the full insurance

[27] Zeldes (1989b) points out how differing marginal tax rates can cause interest r_t to vary across consumers.

scenario, as well as complete markets scenario. This work is well reviewed in Attanasio (1999) and Browning, Hansen and Heckman (1999). Here we present evidence directly related to our discussion of consumption growth above. Two rather effective ways of analyzing failures of the full insurance paradigm fit neatly with our discussion.

One approach to evaluating the full insurance hypothesis is to look directly for evidence that unexpected shocks in income across different groups in the economy lead to differences in consumption patterns (as consistent with (56), which assumes no insurance). This is not a trivial empirical exercise. First, such income shocks have to be identified and measured. Second, there has to be a convincing argument that they would not be correlated with unobservable variables entering marginal utility, or observables such as labor supply (in a nonseparable framework).

Building on the earlier work by Cochrane (1991), Mace (1991), Hayashi, Altonji and Kotlikoff (1996) and Townsend (1994), the study by Attanasio and Davis (1996) presents rigorous and convincing evidence against the full insurance hypothesis using this approach. Low-frequency changes in wages across different education and date-of-birth cohorts are shown to be correlated positively with systematic differences in consumption growth. More recently, Blundell, Pistaferri and Preston (2003) use a combination of the Panel Survey of Income Dynamics (PSID) and the Consumers Expenditure Survey (CES) to investigate insurance of permanent and transitory income shocks at the individual level. They find almost complete insurance to transitory shocks except among lower-income households. They find some insurance to permanent shocks particularly among the younger and higher educated. But they strongly reject the complete insurance model.

The second approach to evaluating full insurance is to assume risk-averse preferences and to model the evolution of idiosyncratic risk terms. In terms of the model (56), this approach examines the relevance of individual risk terms (e.g. σ_{it}^{t-1}) once aggregate risk (σ_{At}^{t-1}) has been allowed for. This is addressed by looking across groups where the conditional variance of wealth shocks is likely to differ over time and to see whether this is reflected in differences in consumption growth. Following earlier work by Dynan (1993), Blundell and Stoker (1999), Caballero (1990) and Skinner (1988), the study by Banks, Blundell and Brugiavini (2001) presents evidence that differential variances of income shocks across date-of-birth cohorts do induce important differences in consumption growth paths.

3.3.2. Aggregation factors and consumption growth

There are two issues. First, if one estimates a model with aggregate data alone, is there likely to be bias in the estimated parameters of interest? Second, will the omission of aggregation bias terms result in spurious inference concerning the presence of excess sensitivity of consumption to transitory income shocks?

With regard to bias, we consider the elasticity of intertemporal substitution ρ, which is normally a focus of studies of aggregate consumption. In Figure 7 we plot the aggre-

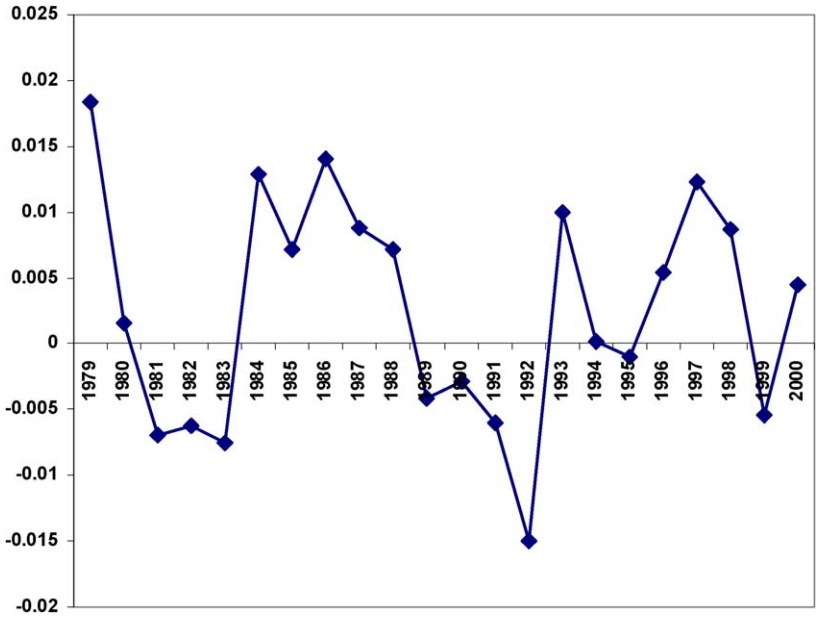

Figure 7. Aggregation factor for consumption growth.

gation factor

$$\Delta \ln E_t(c_{it}) - \Delta E_t(\ln c_{it}) \tag{72}$$

for the sample of married couples from the British FES, used to construct the aggregation factors for demand of Section 2. The figure shows a systematic procyclical variation. We found the correlation coefficient between the real interest series and this factor to be significant. This indicates that there will exist an important aggregation bias in the estimated intertemporal substitution parameter from aggregate consumption data (with a log-linear growth model). This is confirmed in the study by Attanasio and Weber (1993a, 1993b), where aggregate data was constructed from micro survey information.[28] They find an elasticity estimate for aggregate data of around 0.35, and the corresponding micro-level estimates were twice this size.

The study of excess sensitivity involves the use of lagged information as instrumental variables in the estimation of the consumption growth relationship. Omitting aggregation bias terms can invalidate the instruments typically used. For the consumption data used above, we computed the correlation of the aggregation factor with two typically used instrumental variables in consumption growth equations – lagged real interest rates

[28] Attanasio and Weber (1993a, 1993b) also note a strong impact of omitting the cross-section variance of consumption growth.

and lagged aggregate consumption. The estimated correlation coefficient between these series and the omitted bias term was found to be strongly significant.[29]

Together these results suggest that aggregation problems are likely to lead to serious bias in estimated intertemporal substitution parameters and also to exaggerate the presence of excess sensitivity in consumption growth regressions on aggregate data. Attanasio and Browning (1995) investigate this excess sensitivity issue in more detail and find that excess sensitivity still exists at the micro-data level but disappears once controls for age, labour supply variables and demographics are introduced in a flexible way. Moreover, these variables explain why excess sensitivity appears to vary systematically over the cycle.

It is an important finding that evidence of excess sensitivity vanishes once we move to individual data and include observable variables that are likely to impact preferences for the allocation of consumption over time. It has important consequences for our understanding of liquidity constraints and for partial insurance. It has implications for understanding the path of consumption growth over the cycle. It also has implications for the retirement-savings puzzle, or how consumption drops much more at retirement than is predicted by standard consumption growth equations. Banks, Blundell and Tanner (1998) find that once demographics and labor supply variables are allowed to affect the marginal utility of consumption, nearly two thirds of the retirement-savings puzzle disappears.

3.4. Consumption and liquidity constraints

Our previous discussion has focused on heterogeneity in wealth and income risk as it impinges on consumption. We now turn to a discussion of liquidity constraints on consumption, which generate a different kind of aggregation structure. The evidence for liquidity constraints is relatively limited. Most studies of consumption smoothing at the individual level find it difficult to reject the standard model once adequate care is taken in allowing for demographic and labor market interactions; see Attanasio and Weber (1993a, 1993b) and Blundell, Browning and Meghir (1994), for example. Much of the excess sensitivity found in aggregate studies can be attributed to aggregation bias as documented in Attanasio and Weber (1993a, 1993b), Goodfriend (1992) and Pischke (1995). However, there is some evidence that does point to the possibility that a fraction of consumers could be liquidity constrained at particular points in the life-cycle and business cycle. At the micro level some evidence can be found in the studies by Hayashi (1987), Zeldes (1989a), Jappelli (1990), Jappelli and Pagano (1994), Meghir and Weber (1996) and Alessie, Devereux and Weber (1997). As mentioned earlier, the Blundell, Pistaferri and Preston (2003) study shows that the consumption of low-income households in the PSID does react to transitory shocks to income, which suggests that such households do not have access to credit markets to smooth such shocks.

[29] Detailed regression results available on request.

For aggregation, liquidity constraints introduce regime structure into the population. Namely, liquidity-constrained consumers constitute one regime, unconstrained consumers constitute another regime, and aggregate consumption will depend upon the relative distribution across regimes. This structure is particularly relevant for the reaction of consumption growth to increases in current income, since constrained consumers will show a stronger reaction than unconstrained consumers. In this section we discuss these basic issues, and indicate how a model of aggregates can be constructed. Blundell and Stoker (2003) work out the details for aggregate consumption models of this type.

There is some subtlety in considering what population groups are likely to be liquidity constrained. Poor households with a reasonably stable but low expected stream of income may have little reason to borrow. More likely to be constrained are young consumers, who have much human capital but little financial wealth – college students or perhaps poor parents of able children. Such individuals may want to borrow against their future earned incomes but cannot, in part, because their eventual income is higher than others', and the growth of their income with experience is higher. Clearly such consumers will react more than others to shocks in current income and wealth.

We start with the basic consumption model discussed earlier, with permanent and transitory shocks to income. As in (55), the change in current income for consumer i at time t is

$$\Delta \ln y_{it} = \eta_t + \varepsilon_{it} + \Delta u_t + \Delta v_{it}, \tag{73}$$

where $\eta_t + \varepsilon_{it}$ is the permanent component and $\Delta u_t + \Delta v_{it}$ is the transitory component. To keep things simple, we assume that permanent income shocks are not insurable, with log-consumption given as

$$\Delta \ln c_{it} = \rho r_t + (\beta + \varphi r_t)' z_{it} + \eta_t + \varepsilon_{it}, \tag{74}$$

where we assume the precautionary risk terms $(\sigma_{it}^{t-1}, \sigma_{At}^{t-1})$ are included with the z_{it} effects. Note that (74) gives the consumption growth plan $(\rho r_t + (\beta + \varphi r_t)' z_{it})$ as well as how consumption reacts to permanent shocks in income (here $\eta_t + \varepsilon_{it}$).

Liquidity constraints affect the ability of consumers to finance their desired consumption growth path. We follow an approach similar to Zeldes (1989a), where the incidence of liquidity constraints depends on the degree of consumption growth the consumer is trying to finance and the existing stock of assets. In particular, liquidity constraints enter the growth plan only if they are binding in planning period $t-1$, and then the best response will always be to increase consumption growth so as to "jump" back up to the optimal path. If this response is further frustrated by a binding constraint in period t, consumption will simply grow by the amount of resources available.

This response structure is captured by additional terms in Equation (74). Let I_{it} denote the indicator

$$I_{it} = 1[\text{consumer } i \text{ is constrained in period } t-1] \tag{75}$$

and suppose that a consumer who is constrained in period $t-1$ needs to increase consumption growth by m_{it} to return to the optimal growth plan.[30] Then, consumption growth for unconstrained consumers is

$$\Delta \ln c_{it} = \rho r_t + (\beta + \varphi r_t)' z_{it} + I_{it} m_{it} + \eta_t + \varepsilon_{it}. \tag{76}$$

We now model the constraints, as well as consumption growth for constrained consumers. With growth in income of $\Delta \ln y_{it}$, consumer i needs to finance a growth rate of

$$\begin{aligned} &\rho r_t + (\beta + \varphi r_t)' z_{it} + I_{it-1} m_{it} + \eta_t + \varepsilon_{it} - \Delta \ln y_{it} \\ &\quad = \rho r_t + (\beta + \varphi r_t)' z_{it} + I_{it} m_{it} - \Delta u_t - \Delta v_{it} \end{aligned}$$

for consumption at time t to be on the growth plan. To model liquidity constraints at time t, suppose that consumer i faces a borrowing constraint that is associated with a maximum rate of increase of consumption of

$$\gamma + \delta A_{it} + \zeta_{it},$$

where A_{it} is (say) accumulated financial wealth. Consumer i is liquidity constrained in period t, or cannot maintain the consumption growth plan, if

$$\rho r_t + (\beta + \varphi r_t)' z_{it} + I_{it-1} m_{it} - \Delta u_t - \Delta v_{it} > \gamma + \delta A_{it} + \zeta_{it} \tag{77}$$

which we indicate by $I_{it} = 1$, as above. In this case we assume that consumption growth is as large as possible, namely

$$\Delta \ln c_{it} = \Delta \ln y_{it} + \gamma + \delta A_{it} + \zeta_{it}. \tag{78}$$

In terms of permanent and transitory terms of income growth, (78) may be rewritten as

$$\Delta \ln c_{it} = \eta_t + \varepsilon_{it} + \Delta u_t + \Delta v_{it} + \gamma + \delta A_{it} + \zeta_{it}. \tag{79}$$

This is consumption growth for constrained consumers. The constraints have an impact; as consumption growth clearly depends on transitory income shocks and wealth levels.

Aggregate consumption growth will clearly depend on the proportion of consumers who are constrained and the proportion that are not. Consumers who were constrained last period will have a boost in their consumption growth to return to the optimal path. This regime-switching structure is nonlinear in character. Therefore, to model aggregate consumption growth, we would need to specify distributional structure for all the elements that are heterogeneous across the population. We then aggregate over the population of unconstrained individuals with consumption growth (76) and the population of constrained individuals with consumption growth (79). Using log-normality assumptions, we carry out this development in Blundell and Stoker (2003). It is clear how aggregate consumption is affected by transitory income shocks, as well as the distribution of wealth.

[30] Various approaches have been applied to account for the jump term m_{it} in studies of micro-level data. See Zeldes (1989a), Jappelli, Pischke and Souleles (1998), Garcia, Lusardi and Ng (1997), Alessie, Melenberg and Weber (1988), Alessie, Devereux and Weber (1997) and Attanasio and Weber (1993a, 1993b).

3.5. Equilibrium effects

As we mentioned at the start, one use of aggregate consumption equations is to study and understand the evolution of aggregate consumption and saving by themselves. Another important use is in studying equilibrium price and interest rate paths over time. This is an exercise in general equilibrium analysis, and every feature that we have discussed above is relevant – consumer heterogeneity, heterogeneity in income and wealth risks, liquidity constraints, and the distribution of wealth. Further complicating this effort is the dynamic feedback that occurs wherein the level and distribution of wealth evolve as a result of the level and distribution of saving. These difficulties make it very hard to obtain analytical results on equilibrium. Nevertheless, it is extremely important to understand the nature of equilibrium here, including implications on prices and interest rates. We now discuss some recent progress that has been made using calibrated stochastic growth models. A leading example of this effort is provided by Krusell and Smith (1998), although the approach dates from at least Aiyagari (1994a, 1994b) and Heaton and Lucas (1996).

The Krusell–Smith setup has the following features. Consumers are infinitely-lived, with identical (within-period) CRRA preferences, but they are heterogeneous with regard to discount rates. Each consumer has a probability of being unemployed each period, providing transitory, idiosyncratic income shocks. Production arises from a constant returns-to-scale technology in labor and capital, and productivity shocks provide transitory aggregate shocks. Consumers can insure by investing in capital only, so that insurance markets are incomplete, and consumers' capital holdings cannot be negative (liquidity constraint). This setup is rich but in many ways is very simple. Nevertheless, in principle, in order to predict future prices, each consumer must keep track of the evolution of the entire distribution of wealth holdings.

Krusell and Smith's simulations show a rather remarkable simplification to this forecasting problem. For computing equilibrium and for consumer planning, it is only necessary for consumers to keep track of two things, the mean of the wealth distribution and the aggregate productivity shock. Thus there is an informational economy afforded in a similar fashion to a formal aggregation result: once mean wealth is known, the information contained in the distribution of wealth does not appear to improve forecasting very much. This is true even with heterogeneity of many types, including individual and aggregate income shocks (albeit transitory).

The reason for this is clear once the nature of equilibrium is examined. Most consumers, especially those with lowest discount rates, save enough to insure their risk to the point where their propensity to save out of wealth is essentially constant and unaffected by current income or output. Those consumers also account for a large fraction of the wealth. Therefore, saving is essentially a linear function in wealth, and only the mean of wealth matters to how much aggregate saving is done each period. The same is not true of aggregate consumption. There are many low-wealth consumers who become unemployed and encounter liquidity constraints. Their consumption is much more sensitive to current output than that of wealthier consumers. In essence what is happening

here is that the dynamics of the saving process concentrates wealth in the hands of a group that behaves in a homogeneous way, with a constant marginal propensity to save. This (endogenous) simplification allows planning to occur on the basis of mean wealth only.

It is certainly not clear how applicable this finding is beyond the context of this study. This is a computational finding that depends heavily on the specifics of this particular setup.[31] Nonetheless, this form of feedback has some appeal as an explanation of the smooth evolution of wealth distribution, as well as why forecasting equations that fit well are so often much simpler than one would expect from the process that underlies the data. The rich are different (and in this model, the difference makes them rich), but what is important for forecasting is how similar the rich are to one another. With equal saving propensities, it does not matter which group of rich people holds the most wealth.

The study of equilibrium effects and aggregation is in its infancy. Of further note are recent attempts to model differences in micro and macro labor supply elasticities. This includes Chang and Kim (2006) and Rogerson and Wallenius (2007), who incorporate individual decisions at the extensive margin, such as labor participation. In the next section, we discuss labor participation and selection from a partial equilibrium perspective. In any event, we expect the study of equilibrium effects to generate many valuable insights.

4. Wages and labor participation

Our final topic area is the analysis of wages and labor participation. Here the empirical problem is to understand the determinants of wages separately from the determinants of participation. The individual level is that of an individual worker. The economic aggregates to be modeled are aggregate wages and the aggregate participation rate, or one minus the unemployment rate. These statistics are central indicators for macroeconomic policy and for the measurement of economic well-being.

Our analysis is based on a familiar paradigm from labor supply. Potential wages are determined through human capital, and labor participation is determined by comparing potential wages to a reservation wage level. Empirically, there is substantial heterogeneity in the determinants of wages, and substantial heterogeneity in the factors determining labor participation, and both processes are nonlinear. In particular, it is typical to specify wage equations for individuals in log form, and there is much evidence of age and cohort effects in wages and employment. As with demand and consumption, we will need to be concerned with heterogeneity in individual attributes. To keep things as simple as possible, we do not consider forward-looking aspects of employment choice, and so are not concerned with heterogeneity in income and wealth risks.

[31] Carroll (2000) makes a similar argument, with emphasis on the role of precautionary saving. Krusell and Smith (2007) survey recent work, arguing that their original findings are robust to many variations in their framework.

Our primary focus is on heterogeneity in market participation. Aggregate wages depend on the rate of participation, and the important issues involve separation of the wage process from the participation decision. To put it very simply, suppose aggregate wages are increasing through time. Is this because typical wages for workers are increasing? Or, is it because low-wage individuals are becoming unemployed? Do the sources of aggregate wage growth vary other the business cycle? The aggregation problem must be addressed to answer these questions.

We now turn to our basic model of wages that permits us to highlight these effects. We then show the size of these effects for aggregate wages in the UK, a country where there have been large and systematic changes in the composition of the workforce and in hours of work. A more extensive version of this model and the application is given in Blundell, Reed and Stoker (2003). They also summarize derivations of all aggregate equations given below.

4.1. Individual wages and participation

We begin with a model of individual wages in the style of Roy (1951), where wages are based on human capital or skill levels, and any two workers with the same human capital level are paid the same wage. Our framework is consistent with the proportionality hypothesis of Heckman and Sedlacek (1990), where there is no comparative advantage, no sectoral differences in wages for workers with the same human capital level,[32] and the return to human capital is not a function of human capital endowments.

We assume that each worker i possesses a human capital (skill) level of H_i. Suppose human capital is nondifferentiated, in that it commands a single price r_t in each time period t. The wage paid to worker i at time t is

$$w_{ti} = r_t H_i. \tag{80}$$

Human capital H_i is distributed across the population with mean

$$E_t(\ln H_i) = \delta_{js}$$

where δ_{js} is a level that varies with cohort j to which i belongs and education level s of worker i. In other words, the log-wage equation has the additive form

$$\ln w_{it} = \ln r_t + \delta_{js} + \varepsilon_{it} \tag{81}$$

where ε_{it} has mean 0.[33] We will connect δ_{js} to observable attributes below.

To model participation, we assume that reservation wages w_{it}^* are lognormal:

$$\ln w_{it}^* = \alpha \ln B_{it} + \eta_{js} + \zeta_{it}, \tag{82}$$

[32] Heckman and Sedlacek (1985) provide an important generalization of this framework to multiple sectors. See also Heckman and Honoré (1990).

[33] Clearly, there is an indeterminacy in the scaling of r_t and H_i. Therefore, to study r_t, we will normalize r_t for some year $t = 0$ (say to $r_0 = 1$). We could equivalently set one of the δs to zero.

where ζ_{it} has mean 0 and where B_{it} is an exogenous income (welfare benefit) level that varies with individual characteristics and time. Participation occurs if $w_{it} \geqslant w_{it}^*$, or with

$$\ln r_t - \alpha \ln B_{it} + \delta_{js} - \eta_{js} + \varepsilon_{it} - \zeta_{it} \geqslant 0. \tag{83}$$

We represent the participation decision by the indicator $I_{it} = 1[w_{it} \geqslant w_{it}^*]$.

For aggregation over hours of work, it is useful to make one of two assumptions. One is to assume that the distribution of hours is fixed over time. The other is to assume that desired hours h_{it} are chosen by utility maximization, where reservation wages are defined as $h_{it}(w^*) = h_0$ and h_0 is the minimum number of hours available for full-time work. We assume $h_{it}(w)$ is normal for each w, and approximate desired hours by

$$\begin{aligned} h_{it} &= h_0 + \gamma\left(\ln w_{it} - \ln w_{it}^*\right) \\ &= h_0 + \gamma(\ln r_t - \alpha \ln B_{it} + \delta_{js} - \eta_{js} + \varepsilon_{it} - \zeta_{it}). \end{aligned} \tag{84}$$

This is our base-level specification. It is simple to extend this model to allow differentiated human capital, or differential cohort effects due to different labor market experience, which permits a wide range of education/cohort/time effects to be included [c.f. Blundell, Reed and Stoker (2003)]. Because our examples involve log-linear equations and participation (or selection), we summarize the basic framework as

$$\begin{aligned} &\ln w_{it} = \beta_0 + \beta' x_{it} + \varepsilon_{it}, \\ &I_{it} = 1[\alpha_0 + \alpha' z_{it} + \nu_{it} \geqslant 0], \\ &h_{it} = h_0 + \gamma \cdot (\alpha_0 + \alpha' z_{it} + \nu_{it}). \end{aligned} \tag{85}$$

Here, x_{it} denotes education, demographic (cohort, etc.) and time effects, z_{it} includes out-of-work benefit variables, and $I_{it} = 1$ denotes participation. It is clear that the scale of γ is not identified separately from the participation index $\alpha_0 + \alpha' z_{it} + \nu_{it}$; however, we retain γ to distinguish between the fixed hours case $\gamma = 0$ and the variable hours case $\gamma \neq 0$.

Our notation distinguishes two types of individual heterogeneity in (85). The variables x_{it} and z_{it} are observable at the individual level, while ε_{it} and ν_{it} are unobservable. Analysis of data on wages and participation at the individual level requires assumptions on the distribution of those unobservable elements, a process familiar from the literature on labor supply and selection bias. We now review some standard selection formulae here for later comparison with the aggregate formulations. Start with the assumption that the unobserved elements are normally distributed

$$\begin{pmatrix} \varepsilon_{it} \\ \nu_{it} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon\nu} \\ \sigma_{\varepsilon\nu} & \sigma_\nu^2 \end{pmatrix}\right). \tag{86}$$

This allows us to apply some well-known selection formulae (given in virtually every textbook of econometrics). The micro participation regression, or the proportion of par-

ticipants given x_{it} and z_{it}, is a probit model:

$$E_t[I|x_{it}, z_{it}] = \Phi\left[\frac{\alpha_0 + \alpha' z_{it}}{\sigma_v^2}\right]. \tag{87}$$

The micro log-wage regression for participants is

$$E_t[\ln w_{it}|I_{it} = 1,\ x_{it}, z_{it}] = \beta_0 + \beta' x_{it} + \frac{\sigma_{\varepsilon v}}{\sigma_v}\lambda\left[\frac{\alpha_0 + \alpha' z_{it}}{\sigma_v^2}\right] \tag{88}$$

reflecting the typical (Heckman-style) selection term, which adjusts the log-wage equation to the group of participating workers.[34]

4.2. Aggregate wages and employment

The aggregate of interest is average hourly earnings, where aggregation occurs over all workers, namely

$$\overline{w}_t = \frac{\sum_{i\in(I=1)} h_{it} w_{it}}{\sum_{i\in(I=1)} h_{it}} = \sum_{i\in(I=1)} \mu_{it} w_{it}, \tag{89}$$

where $i \in (I = 1)$ denotes a participant (worker), $h_{it} w_{it}$ is the earnings of individual i in period t, and μ_{it} are the hours-weights

$$\mu_{it} = \frac{h_{it}}{\sum_{i\in(I=1)} h_{it}}.$$

Modelling the aggregate wage (89) requires dealing with log-nonlinearity of the basic wage equation, dealing with participation and dealing with the hours-weighting. All of these features require that distributional assumptions be made for (observable) individual heterogeneity. In particular, we make the following normality assumption for x_{it} and z_{it}:

$$\begin{pmatrix} \beta_0 + \beta' x_{it} \\ \alpha_0 + \alpha' z_{it} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta_0 + \beta' E(x_{it}) \\ \alpha_0 + \alpha' E(z_{it}) \end{pmatrix}, \begin{pmatrix} \beta' \Sigma_{xx} \beta & \alpha' \Sigma_{xz}' \beta \\ \beta' \Sigma_{xz} \alpha & \alpha' \Sigma_{zz} \alpha \end{pmatrix}\right) \tag{90}$$

or that the indices determining log-wage and employment are joint normally distributed.[35]

We now discuss some aggregate analog of the micro regression equations, and then our final equation for the aggregate wage. The aggregate participation (employment)

[34] Here $\Phi(\cdot)$ is the normal cumulative distribution function, and $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$, where $\phi(\cdot)$ is the normal density function.

[35] Assuming that the linear indices are normal is much weaker than assuming that x_{it} and z_{it} are themselves joint multivariate normal. Such a strong structure would eliminate many important regressors, such as qualitative variables.

rate is

$$E_t[I] = \Phi\left[\frac{\alpha_0 + \alpha' E(z_{it})}{\sqrt{\alpha' \Sigma_{zz}\alpha + \sigma_\nu^2}}\right], \tag{91}$$

using a formula originally due to McFadden and Reid (1975). Aggregate participation has the same form as the micro participation regression (87) with z_{it} replaced by $E(z_{it})$ and the spread parameter σ_ν^2 replaced by the larger value $\sqrt{\alpha' \Sigma_{zz}\alpha + \sigma_\nu^2}$, reflecting the influence of heterogeneity in the individual attributes that affect the participation decision. The mean of log-wages for participating (employed) workers is

$$E_t[\ln w_{it} | I_{it} = 1] = \beta_0 + \beta' E_t(x_{it} | I = 1) + \frac{\sigma_{\varepsilon\nu}}{\sqrt{\alpha' \Sigma_{zz}\alpha + \sigma_\nu^2}} \lambda\left[\frac{\alpha_0 + \alpha' E_t(z_{it})}{\sqrt{\alpha' \Sigma_{zz}\alpha + \sigma_\nu^2}}\right], \tag{92}$$

using a formula originally derived by MaCurdy (1987). This matches the micro log-wage regression (88) with x_{it} replaced by $E_t(x_{it}|I = 1)$, z_{it} replaced by $E_t(z_{it})$ and the spread parameter changed from σ_ν^2 to $\sqrt{\alpha' \Sigma_{zz}\alpha + \sigma_\nu^2}$. This is an interesting result, but does not deliver an equation for the aggregate wage $\overline{w}_t$.

Blundell, Reed and Stoker (2003) derive such an equation. The aggregate wage is given as

$$\ln \overline{w}_t = \ln \frac{E_t[h_{it} w_{it} | I_{it} = 1]}{E_t[h_{it} | I_{it} = 1]} = \beta_0 + \beta' E_t(x_{it}) + [\Omega_t + \Psi_t + \Lambda_t], \tag{93}$$

where the aggregation bias is comprised of a spread term

$$\Omega_t = \frac{1}{2}[\beta' \Sigma_{xx} \beta + \sigma_\varepsilon^2], \tag{94}$$

plus two terms Ψ_t and Λ_t, which represent separate sources of bias but have very complicated expressions.[36]

What these terms represent can be seen most easily by the following construction. Begin with the individual wage equation evaluated at mean attributes, $\beta_0 + \beta' E_t(x_{it}) = E_t(\ln w_{it})$, or overall mean log-wage. Adding Ω_t adjusts for log-nonlinearity, as

$$\ln E_t(w_{it}) = E_t(\ln w_{it}) + \Omega_t.$$

[36] In particular, we have

$$\Psi_t \equiv \ln\left\{\Phi\left[\frac{\alpha_0 + \alpha' E(z_{it}) + \beta' \Sigma_{xz}\alpha + \sigma_{\varepsilon\nu}}{\sqrt{\alpha' \Sigma_{zz}\alpha + \sigma_\nu^2}}\right] \Big/ \Phi\left[\frac{\alpha_0 + \alpha' E(z_{it})}{\sqrt{\alpha' \Sigma_{zz}\alpha + \sigma_\nu^2}}\right]\right\},$$

$$\Lambda_t \equiv \ln\left[\frac{h_0 + \gamma\alpha_0 + \gamma\alpha' E(z_{it}) + \gamma\beta' \Sigma_{xz}\alpha + \gamma\sigma_{\varepsilon\nu} + \gamma\sqrt{\alpha' \Sigma_{zz}\alpha + \sigma_\nu^2} \cdot \lambda^a_{\sigma_{\varepsilon\nu},t}}{h_0 + \gamma\alpha_0 + \gamma\alpha' E(z_{it}) + \gamma\sqrt{\alpha' \Sigma_{zz}\alpha + \sigma_\nu^2} \cdot \lambda^a_t}\right].$$

Adding Ψ_t adjusts for participation, as

$$\ln E_t[w_{it}|I_{it} = 1] = \ln E_t(w_{it}) + \Psi_t. \tag{95}$$

Finally, adding Λ_t adjusts for hours-weighting, as

$$\ln \overline{w}_t = \ln E_t[w_{it}|I_{it} = 1] + \Lambda_t = E_t(\ln w_{it}) + \Omega_t + \Psi_t + \Lambda_t. \tag{96}$$

Thus, the bias expressions are complicated but the roles of Ω_t, Ψ_t and Λ_t are clear. In words, the term Ω_t captures the variance of returns, observable and unobservable. The term Ψ_t reflects composition changes within the selected sample of workers from which measured wages are recorded. The term Λ_t reflects changes in the composition of hours and depends on the size of the covariance between wages and hours.

The formulation (93) of the log aggregate wage $\ln \overline{w}_t$ thus captures four important sources of variation. First, aggregate wages increase if the distribution of log-wages shifts to the right, which is the typical "well-being" interpretation of aggregate wage movements.[37] This source is reflected by the mean $\beta_0 + \beta' E(x_{it})$ of log-wages. Second, because individual wages are given in log form, aggregate wages will increase with increased spread of the log-wage distribution, as reflected by the heterogeneity term Ω_t. Third, aggregate wages will increase if the benefit threshold increases, causing more lower-wage individuals to decide not to participate. This is reflected in the participation term Ψ_t. Fourth, aggregate wages will increase if the hours of higher-wage individuals increase relative to lower-wage individuals, which is captured by the hours adjustment term Λ_t. The aggregate model (93) permits estimation of these separate effects.

This framework could be relaxed in many ways. We can allow all variance terms to be time varying, as well as many of the basic behavioral parameters. If the normality assumption on the overall log-wage and participation index is not accurate for the whole population, the population can be segmented, with separate aggregate equations developed for each segment. These variations, among others, are discussed in Blundell, Reed and Stoker (2003).

4.3. Empirical analysis of British wages

The different sources of aggregate wage variation bear directly on the issue of whether aggregate wages are procyclical or not. In particular, the participation effect works counter to a normal cyclical variation of aggregate wages – decreases in participation can lead to aggregate wage increases when there is essentially no change in individual

[37] Comparing $\ln \bar{w}_t$ to mean log-wage $E_t(\ln w_{it})$ is in line with the tradition of measuring "returns" from coefficients in log-wage equations estimated with individual data; c.f. Solon, Barksy and Parker (1994). Other comparisons are possible, and some may be preferable on economic grounds. For instance, if aggregate production in the economy has total human capital $(\sum_i H_i)$ as an input, then the appropriate price for that input is r_t, so one might want to compare $\ln \overline{w}_t$ to $\ln r_t$ for a more effective interpretation. In any case, it is useful to point out that if $E(\ln H_i)$ is constant over time, then comparing $\ln r_t$ to $\ln \overline{w}_t$ is the same as comparing $E_t(\ln w_{it})$ to $\ln \overline{w}_t$.

wage levels or distribution. We now turn to an analysis of British wages that shows these features.

Our microeconomic data are again taken from the UK Family Expenditure Survey (FES), for the years 1978 to 1996. The FES is a repeated continuous cross-section survey which contains consistently defined micro data on wages, hours of work, employment status and education for each year since 1978. Our sample consists of all men aged between 19 and 59 (inclusive).[38] The participating group consists of employees; the nonparticipating group includes individuals categorized as searching for work as well as the unoccupied. The hours measure for employees in FES is defined as usual weekly hours including usual overtime hours, and weekly earnings includes overtime pay. We divide nominal weekly earnings by weekly hours to construct an hourly wage measure, which is deflated by the quarterly UK retail price index to obtain real hourly wages.

Individual attributes include education level and cohort effects. Individuals are classified into three educational groups: those who left full-time education at age 16 or lower, those who left aged 17 or 18, and those who left aged 19 or over. Dummy variables capture effects of five date-of-birth cohorts (b.1919–1934, b.1935–1944, b.1945–1954, b.1955–1964 and b.1965–1977). We include various trend variables to account for a common business-cycle effect. Finally, our measure of benefit income (income at zero hours) is constructed for each individual as described in Blundell, Reed and Stoker (2003). After making the sample selections described above, our sample contains 40,988 observations, of which 33,658 are employed, or 82.1% of the total sample.

4.3.1. Real wages and employment

Figure 8 shows log average wages in Britain from 1978 to 1996. These show a strong trend increase over the whole period. The trend appears for more disaggregate groups. Blundell, Reed and Stoker (2003) present a more detailed breakdown by cohort, region and education group, and show that the trend holds widely, including for the least-educated group.

Figure 9 shows the overall male labor employment rate for the same period. Clearly there has been a large fall in the participation rate of men. Figure 10 presents the employment rate for those with low education. For this group, there is a continued and much steeper decline in employment. This period also included two deep recessions in which there have been large fluctuations in male employment.

Considering Figures 8–10 together, one can understand the basic importance of sorting out wage growth at the individual level from changes in participation. The strong trend of aggregate wages is suggestive of great progress at increasing the well-being of

[38] We exclude individuals classified as self-employed. This could introduce some composition bias, given that a significant number of workers moved into self-employment in the 1980s. However, given that we have no data on hours and relatively poor data on earnings for this group, there is little alternative but to exclude them. They are also typically excluded in aggregate figures.

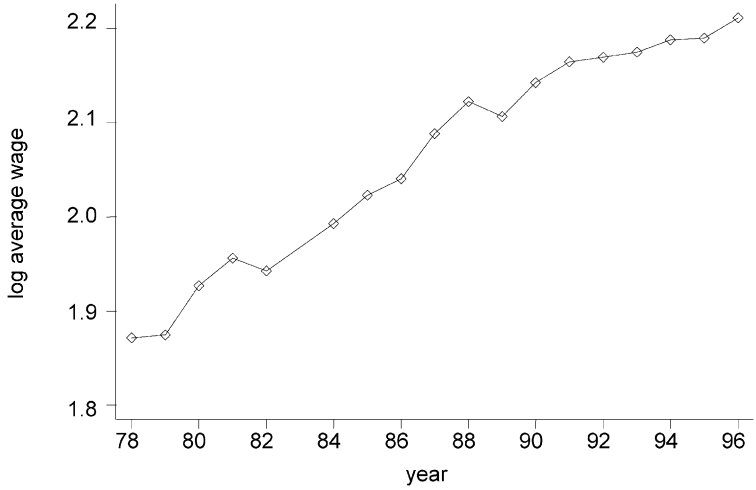

Figure 8. Male hourly wages.

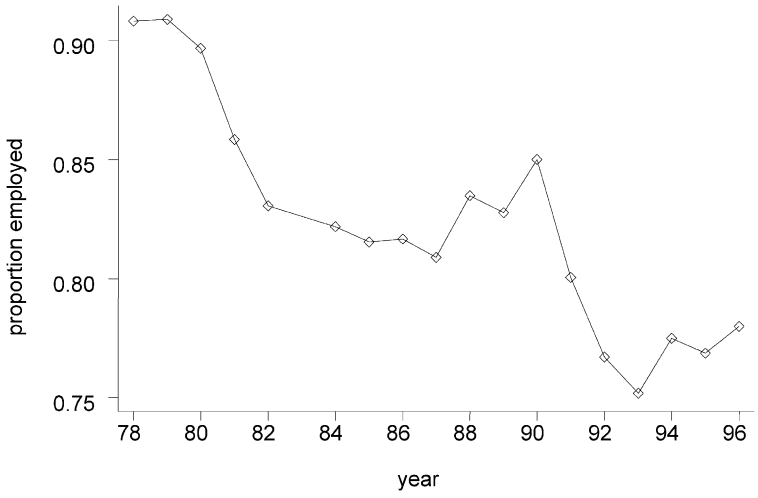

Figure 9. Male employment rate.

laborers in general.[39] However, great increases in unemployment are likely associated with unemployment of workers with lowest wages, or workers from the poorest groups. It is very important to understand how much of the upward trend in average wages is due to the elimination of low-wage earners from employment.

[39] In fact, such a conclusion has been trumpeted by British newspapers.

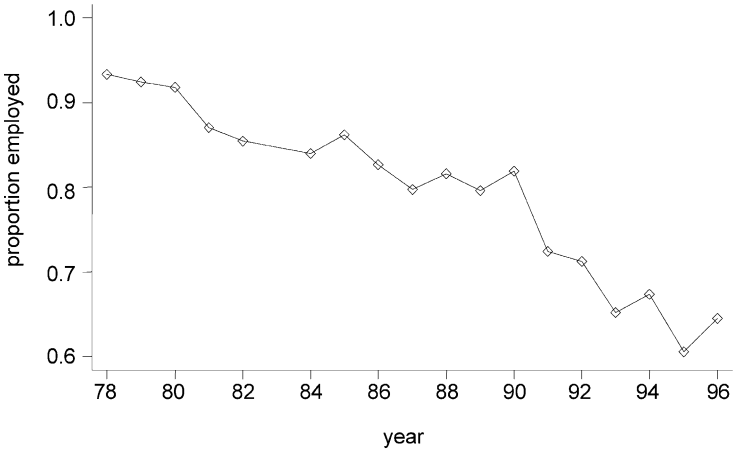

Figure 10. Employment rate for the low-educated.

There have also been well-documented changes in real benefit income over time and across different groups of individuals. While it is unlikely that variation in real value of benefit income relative to real earnings can explain all of the variation in participation rates, the changes in real benefits act as an important "instrumental variable" for separating participation decisions from determinants of wages. Again, to the extent that changes in benefit income have discouraged (or encouraged) participation, it is essential to learn the size of this impact relative to the other factors driving changes in wages.

4.3.2. Aggregation results

The Blundell, Reed and Stoker (2003) study considers a number of possible specifications for our individual-level wage equations which relate to the various specifications. In the simplest of our specifications, the full proportionality hypothesis is imposed on the (nondifferentiated) human capital model, together with trend terms to reflect the business-cycle effects on skill price. This specification was strongly rejected by the data. The preferred model had full interactions of cohort, trend, region and education. These additional variables could reflect many differences in minimum educational standards across cohorts such as the systematic raising of the minimum school-leaving age over the postwar period in the UK. The prices of different (education-level) skills are allowed to evolve in different ways, by including an interaction between high education and the trend terms. These coefficients are marginally significant and show an increasing trend among groups with higher levels of human capital. The impact of adjusting for

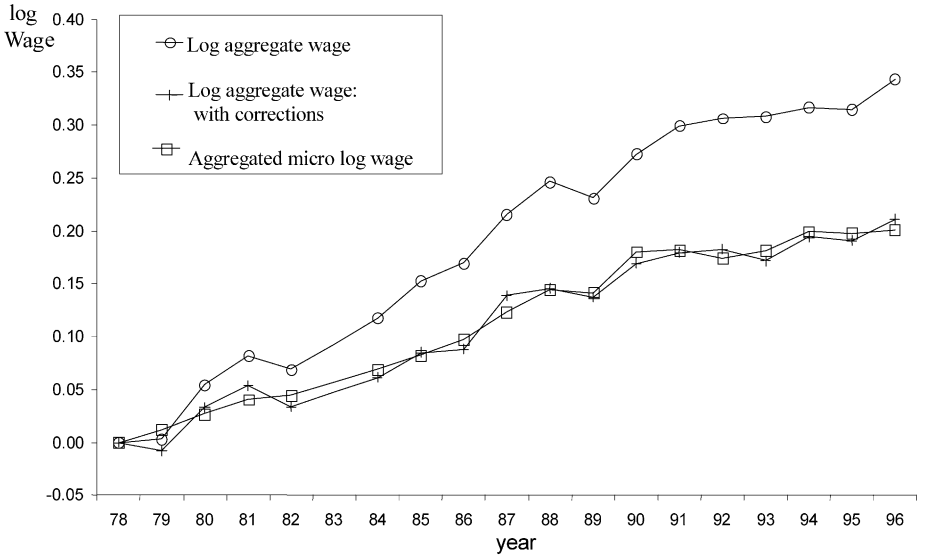

Figure 11. Hourly wage growth.

participation is very important.[40] To see the impact of these results on aggregate wages, we turn to graphical analysis.

Figure 11 displays the (raw) log aggregate wage, the log aggregate wage minus the estimated aggregation bias terms, and the mean of the log wage from the selectivity-adjusted micro model. We have plotted the return from a common point at the start of the time-series rebased to zero for 1978, to highlight the changes in trend growth in wages indicated by our corrections. There is a clear downward shift in the trend, and an increased cyclical component in wage growth shown by both the corrected aggregate series and the estimated micro model.

This procedure is repeated for the lower-education group in Blundell, Reed and Stoker (2003). Several features of this analysis are worth mentioning here. For instance, even the direction of movement of the uncorrected log aggregate wage does not always mirror that of the mean micro log wage. There is a reasonably close correspondence between the two in the 1984–1988 period, but the 1990–1993 period is different. In 1990–1993, log aggregate wages are increasing, but the mean micro log wage (and the corrected aggregate wage) is decreasing – precisely the period where there is a big decline in participation. What is remarkable is that the aggregate data show reasonable growth in real wages, but such growth is virtually absent from the corrected series. We are left with a much more cyclical profile of wages.

[40] Blundell, Reed and Stoker (2003) examined the impact of our normality assumptions by estimating with semiparametric methods. The estimated wage coefficients were hardly affected by this generalization.

If the model is exactly correct, the results from aggregating the selectivity-adjusted micro model estimates should match the corrected aggregate series. They show a close correspondence in Figure 11, and a similar close correspondence is noted by Blundell, Reed and Stoker (2003) for more disaggregated groups.[41] In any case, we view the correspondence between the corrected log aggregate micro wage and the mean micro log wage as striking validation of the framework. This model specification that provides a good and parsimonious specification of the evolution of log real wages also seems to work well in terms of the specification of aggregation factors.

5. Conclusion

Macroeconomics is one of the most important fields of economics. It has perhaps the grandest goal of all economic study, which is to advise policymakers who are trying to improve the economic well-being of entire populations of people. In the mid-twentieth century, say 1940 to 1970, macroeconomics had an orientation toward its role much like an oracle giving advice while peering down from the top of a mountain. That is, while economists could see people making detailed decisions about buying products, investing their wealth, choosing jobs or career paths, etc., macroeconomic models were extremely simple. For instance, describing the aggregate consumption of an entire economy could be done by taking into account just a few variables: aggregate income, lagged aggregate consumption, etc. Such equations often fit aggregate data extremely well. Unfortunately, such models could not predict future aggregate variables with sufficient precision to dictate optimal policies. Even with great statistical fit, there was too much uncertainty as to what the underlying processes were that drove the aggregate data, and for policy prescriptions it is crucial to know something about those processes.[42]

What economists could get a handle on was how rational individuals and firms would behave in various economic environments. Problems such as how to allocate one's budget, how much to save and invest, or whether to work hard or not so hard, are sufficiently familiar that their essence could be captured with some mathematics, and

[41] To get an idea of the precision of these results, Blundell, Reed and Stoker (2003) present bootstrap 95% confidence bands for the corrected log wage estimates for various groups. These plots show that the micro model prediction and the corrections to the log aggregate wage are both quite tightly estimated. In all cases, the micro model prediction and the corrections to the aggregate wage plot are significantly different from the raw aggregate wage measure and not significantly different from each other. This gives us confidence that we have identified compositional biases in the measured real wage with a reasonable degree of precision.

[42] There are many stories told in the economics profession about what giants of our field thought were the greatest contributions to social science. In this spirit, we relate the following. In the mid-1980s, one of the authors asked Paul Samuelson what he felt was the greatest failure in economics. Without hesitation, his answer was "macroeconomics and econometrics". The reason for this is that there had been an enormous anticipation in the 40s, 50s and 60s that simple empirical macroeconomic models would, in fact, be accurate enough to allow real economies to be guided and controlled, much like an automobile or a spacecraft. That this turned out to not be possible was a source of great disappointment.

economists could describe and prescribe optimal reactions. Economists could settle how someone being really smart and clear-headed would behave. Notwithstanding the anomalies pointed out recently by behavioral economists, the predictive power of economics rests on the notion that people facing a familiar situation will behave in their interests. Foolish, self-destructive or purely random behavior will not be repeated once it is consciously seen to be less good than another course. The transformation of economic analysis by mathematics occurred through the systematic understanding of rational and learning behavior by individuals and firms, and the overall implications of that for market interactions.

The merging of these two bodies of thought – macroeconomics and optimal behavior of individuals – is among the greatest developments of economics in the last half century. This advance has been recognized by Nobel prizes to Lucas, Kydland and Prescott, and one should expect more prizes to be awarded to other important developers. Previous "schools of thought" have been replaced by groups differentiated by how they settle the tradeoff between realism and strict adherence to optimal economic behavior. The specification of macroeconomic models, the judgment of whether they are sensible, and the understanding of the impacts of economic policy are now more systematic because of their embedding in the rules of optimal individual behavior.

The trouble is, this embedding cannot be right without taking account of aggregation. A one-person or five-person economy is just not realistic. One can simulate a model with a few actors and pretend that it is realistic, but there is nothing in casual observation or empirical data or economic theory that suggests that such a stance is valid. There is much to be learned from rational individual behavior, but there must be an explicit bridge to economic aggregates because real people and their situations are so very heterogeneous. Aggregation is essential, because heterogeneity is a pervasive and indisputable fact of life.

In this chapter, we have covered recent work on aggregation problems in a style that we hope is useful to empirical economists. Our orientation has been to highlight the importance of different types of individual heterogeneity: in particular, heterogeneity in tastes and reaction, heterogeneity in market participation, and heterogeneity in uninsurable risks. Our approach has been practical; we have covered recent advances in econometric modeling that address issues in aggregation, by considering explicit models at the individual level and among economic aggregates.

We have covered a wide range of ideas. First, we have detailed the main approach for incorporating distributional information into aggregate relationships, namely exact aggregation models, in the context of how that approach has been applied to the analysis of consumer demands. Second, we have shown how one can incorporate basic nonlinearity, insurance and dynamic elements, in our coverage of aggregate consumption based on CRRA preferences. Third, we have shown how to account for compositional heterogeneity, in our coverage of labor participation and wages. The latter two topics required explicit assumptions on the distribution of individual heterogeneity, and we have based our solutions on normal and lognormal assumptions on individual heterogeneity. While these distributional restrictions are specific, they do permit explicit

formulations of the aggregate relationships of interest to be derived, and those formulations capture both location and spread (mean and variance) of the underlying elements of individual heterogeneity. We view our solutions in these cases as representative and clear, and good starting points for empirical modeling in the respective areas.

Whether one dates the beginning of the study of aggregation problems from the 1940s, 1930s or perhaps earlier, one can at best describe progress toward solutions as slow. Aggregation problems are among the most difficult problems faced in either the theoretical or empirical study of economics. Heterogeneity across individuals is extremely extensive and its impact is not obviously simplified or lessened by the existence of economic interaction via markets or other institutions. The conditions under which one can ignore a great deal of the evidence of individual heterogeneity are so severe as to make them patently unrealistic. There is no quick, easy or obvious fix to dealing with aggregation problems in general.

Yet we see the situation as hopeful and changing, and offer the solutions discussed in this chapter as evidence of that change. The sources of this change are two-fold, and it is worth pointing them out as well as pointing out how both are necessary.

The first source of change is the increasing availability of data on individuals observed over sequential time periods. To address questions of what kinds of individual heterogeneity are important for aggregate relationships, one must assess what kinds of heterogeneity are relevant to individual behavior for the problem at hand, and assess how much the distributions of the relevant heterogeneity vary over time. To the extent that this heterogeneity reflects differences in unexpected shocks to individual agents, the mechanisms that are available to individuals to insure against such shocks will have a strong bearing on the form of the aggregate relationship.

While we have advanced the idea of using aggregation factors (derived from time-series of individual data) to summarize the impacts of aggregation, the specific method one uses is less important than the ability to use all available types of information to study economic relationships. That is, it is important to study any relationship among economic aggregates with individual data as well as aggregate data, to get as complete a picture as possible of the underlying structure. Even though modeling assumptions will always be necessary to develop explicit formulations of aggregate relationships, testing those assumptions is extremely important, and is not possible without extensive individual data over sequential time periods. Our view is that the prospects for meaningful advance continue to brighten, as the data situation with regard to individual behavior and aggregate economic variables will continue to improve.

The second source of change in studying aggregation problems is the recent, rapid rise in computing power. Realistic accommodation of individual heterogeneity typically requires extensive behavioral models, let alone combinations of individual models with aggregate relationships. Within the last twenty five years (or the professional lives of both authors), there have been dramatic changes in the ability to implement realistic models. Before this, it was extremely difficult to implement models that are necessary for understanding impacts of individual heterogeneity in aggregation.

Aggregation problems remain among the most vexing in all of applied economics. While they have not become less difficult in the past decade, it has become possible to study aggregation problems in a meaningful way. As such, there are many reasons to be optimistic about the prospects for steady progress on aggregation problems in the future. The practice of ignoring or closeting aggregation problems as "just too hard" is no longer appropriate.

References

Aiyagari, S.R. (1994a). "Uninsured idiosyncratic risk and aggregate savings". Quarterly Journal of Economics 104 (2), 219–240.

Aiyagari, S.R. (1994b). "Co-existence of a representative agent type equilibrium with a nonrepresentative agent type equilibrium". Journal of Economic Theory 57 (1), 230–236.

Alessie, R., Devereux, M., Weber, G. (1997). "Intertemporal consumption, durables and liquidity constraints: A cohort analysis". European Economic Review 41, 37–60.

Alessie, R., Melenberg, B., Weber, G. (1988). "Consumption, leisure and earnings-related liquidity constraints: A note". Economics Letters 27, 101–104.

Atkeson, A., Ogaki, M. (1996). "Wealth-varying intertemporal elasticities of substitution: Evidence from panel and aggregate data". Journal of Monetary Economics 38, 507–534.

Atkinson, A.B., Gomulka, J., Stern, N.H. (1990). "Spending on alcohol: Evidence from the Family Expenditure Survey 1970–1983". Economic Journal 100, 808–827.

Attanasio, O.P. (1999). "Consumption". In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics. North-Holland, Amsterdam.

Attanasio, O.P., Browning, M. (1995). "Consumption over the life cycle and over the business cycle". American Economic Review 85 (5), 1118–1137.

Attanasio, O.P., Davis, S.J. (1996). "Relative wage movements and the distribution of consumption". Journal of Political Economy 104, 1227–1262.

Attanasio, O.P., Weber, G. (1993a). "Is consumption growth consistent with intertemporal optimization? Evidence from the Consumer Expenditure Survey". Journal of Political Economy 103 (6), 1121–1157.

Attanasio, O.P., Weber, G. (1993b). "Consumption growth, the interest rate and aggregation". Review of Economic Studies 60, 631–649.

Banks, J., Blundell, R.W., Brugiavini, A. (2001). "Risk pooling, precautionary saving and consumption growth". Review of Economic Studies 68(4) (237), 757–779.

Banks, J., Blundell, R.W., Lewbel, A. (1997). "Quadratic Engel curves, indirect tax reform and welfare measurement". Review of Economics and Statistics 79 (4), 527–539.

Banks, J., Blundell, R.W., Tanner, S. (1998). "Is there a retirement-savings puzzle?". American Economic Review 88 (4), 769–788 (September).

Barten, A.P. (1964). "Family composition, prices and expenditure patterns". In: Hart, P.E., Mills, F., Whitaker, J.K. (Eds.), Econometric Analysis for National Economic Planning. Butterworths, London.

Berry, S., Levinsohn, J., Pakes, A. (2004). "Differentiated products demand systems from a combination of micro and macro data: The new car market". Journal of Political Economy 112, 68–105.

Bierens, H.J., Pott-Buter, H.A. (1990). "Specification of household Engel curves by nonparametric regression". Econometric Reviews 9, 123–184.

Blackorby, C., Primont, D., Russell, R.R. (1978). Duality, Separability and Functional Structure: Theory and Economic Applications. North-Holland, Amsterdam.

Blundell, R.W. (1988). "Consumer behaviour: Theory and empirical evidence". Economic Journal 98, 16–65.

Blundell, R.W., Browning, M., Crawford, I. (2003). "Nonparametric Engel curves and revealed preference". Econometrica 71 (1), 205–240 (January).

Blundell, R.W., Browning, M., Meghir, C. (1994). "Consumer demand and life-cycle allocation of household expenditures". Review of Economic Studies 61, 57–80.
Blundell, R.W., Duncan, A., Pendakur, K. (1998). "Semiparametric estimation and consumer demand". Journal of Applied Econometrics 13, 435–461.
Blundell, R.W., Pashardes, P., Weber, G. (1993). "What do we learn about consumer demand patterns from micro-data?". American Economic Review 83 (3), 570–597 (June).
Blundell, R., Pistaferri, L., Preston, I. (2003). "Consumption inequality and partial insurance". UCL Discussion Paper No. 03/06, October.
Blundell, R.W., Reed, H., Stoker, T. (2003). "Interpreting aggregate wage growth". American Economic Review 93 (4), 1114–1131 (September).
Blundell, R.W., Robin, J.-M. (2000). "Latent separability: Grouping goods without weak separability". Econometrica 68 (1), 53–84.
Blundell, R.W., Stoker, T. (1999). "Consumption and the timing of income risk". European Economic Review 43 (3), 475–507 (March).
Blundell, R., Stoker, T. (2003). "Models of aggregate economic relationships that account for heterogeneity". Preliminary manuscript. December.
Brown, D.J., Matzkin, R.L. (1996). "Testable restrictions on the equilibrium manifold". Econometrica 64 (6), 1249–1262.
Browning, M. (1992). "Children and household economic behavior". Journal of Economic Literature 30, 1434–1475.
Browning, M., Hansen, L., Heckman, J.J. (1999). "Micro data and general equilibrium models". In: Taylor, J., Woodford, M. (Eds.), Handbook of Macroeconomics. Elsevier, Amsterdam, pp. 543–633. Chapter 8.
Caballero, R.J. (1990). "Consumption puzzles and precautionary savings". Journal of Monetary Economics 25, 113–136 (January).
Carroll, C.D. (2000). "Requiem for the representative consumer? Aggregate implications of microeconomic consumption behavior". American Economic Review Papers and Proceedings 90, 110–115.
Chang, Y., Kim, S. (2006). "From individual to aggregate labor supply: A quantitative analysis based on a heterogeneous agent macroeconomy". International Economic Review 47, 1–27.
Chiappori, P.-A. (1988). "Rational household labor supply". Econometrica 56, 63–90.
Chiappori, P.-A. (1994). "The collective approach to household behaviour". In: Blundell, R.W., Preston, I., Walker, I. (Eds.), The Measurement of Household Welfare. Cambridge University Press, Cambridge, pp. 70–86.
Cochrane, J.H. (1991). "A simple test of consumption insurance". Journal of Political Economy 99, 957–976.
Deaton, A., Paxson, C. (1994). "Intertemporal choice and inequality". Journal of Political Economy 102, 4437–4467.
Deaton, A.S., Muellbauer, J. (1980a). "An almost ideal demand system". American Economic Review 70, 312–326.
Deaton, A.S., Muellbauer, J. (1980b). Economics and Consumer Behavior. Cambridge University Press, Cambridge.
Diewert, W.E. (1977). "Generalized Slutsky conditions for aggregate demand functions". Journal of Economic Theory 15, 353–362.
Dynan, K.E. (1993). "How prudent are consumers?". Journal of Political Economy 101, 1104–1113.
Forni, M., Lippi, M. (1997). Aggregation and the Microfoundations of Dynamic Macroeconomics. Clarendon Press, Oxford.
Garcia, R., Lusardi, A., Ng, S. (1997). "Excess sensitivity and asymmetries in consumption". Journal of Money Credit and Banking 29, 154–176.
Goodfriend, M. (1992). "Information aggregation bias". American Economic Review 82, 508–519.
Gorman, W.M. (Terence) (1953). "Community preference fields". Econometrica 21, 63–80.
Gorman, W.M. (Terence) (1981). "Some Engel curves". In: Deaton, A.S. (Ed.), Essays in the Theory and Measurement of Consumer Behavior. Cambridge University Press, Cambridge.
Grandmont, J.-M. (1992). "Transformations of the commodity space: Behavioral heterogeneity and the aggregation problem". Journal of Economic Theory 57, 1–35.

Granger, C.W.J. (1980). "Long-memory relationships and the aggregation of dynamic models". Journal of Econometrics 14, 227–238.

Granger, C.W.J. (1987). "Implications of aggregation with common factors". Econometric Theory 3, 208–222.

Granger, C.W.J. (1990). "Aggregation of time series variables: A survey". In: Barker, T.S., Pesaran, M.H. (Eds.), Disaggregation in Econometric Modeling. Routledge, London.

Hall, R.E. (1978). "Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence". Journal of Political Economy 86, 971–987.

Härdle, W., Hildenbrand, W., Jerison, M. (1991). "Empirical evidence for the law of demand". Econometrica 59, 1525–1550.

Hausman, J.A., Newey, W.K., Powell, J.L. (1994). "Nonlinear errors-in-variables: Estimation of some Engel curves". Journal of Econometrics 65, 205–233.

Hayashi, F. (1987). "Tests for liquidity constraints: A critical survey". In: Truman, B. (Ed.), In: Advances in Econometrics: Fifth World Congress, vol. 2. Cambridge University Press, Cambridge.

Hayashi, F., Altonji, J., Kotlikoff, L. (1996). "Risk-sharing between and within families". Econometrica 64, 261–294.

Heaton, J., Lucas, D. (1996). "Evaluating the effects of incomplete markets on risk sharing and asset pricing". Journal of Political Economy 104, 443–487.

Heckman, J.J., Honoré, B.E. (1990). "The empirical content of the Roy model". Econometrica 58, 1121–1149.

Heckman, J.J., Sedlacek, G. (1985). "Heterogeneity, aggregation and market wage functions: An empirical model of self-selection in the labor market". Journal of Political Economy 93, 1077–1125.

Heckman, J.J., Sedlacek, G. (1990). "Self-selection and the distribution of hourly wages". Journal of Labor Economics 8, S329–S363.

Hildenbrand, W. (1994). Market Demand: Theory and Empirical Evidence. Princeton University Press, Princeton.

Hildenbrand, W. (1998). "How relevant are specifications of behavioural relations on the micro-level for modelling the time path of population aggregates?". European Economic Review 42, 437–458.

Hildenbrand, W., Kneip, A. (1993). "Family expenditure data, heteroscedasticity and the law of demand". Ricerche Economiche 47, 137–165.

Imbens, G.W., Lancaster, T. (1994). "Combining micro and macro data in microeconometric models". Review of Economic Studies 61, 655–680.

Jappelli, T. (1990). "Who is credit constrained in the US economy?". Quarterly Journal of Economics 105, 219–234.

Jappelli, T., Pagano, M. (1994). "Saving, growth and liquidity constraints". Quarterly Journal of Economics 109, 83–110.

Jappelli, T., Pischke, J.-S., Souleles, N. (1998). "Testing for liquidity constraints in Euler equations with complementary data sources". Review of Economics and Statistics 80 (2), 251–262.

Jorgenson, D.W., Lau, L.J., Stoker, T.M. (1980). "Welfare comparisons and exact aggregation". American Economic Review 70, 268–272.

Jorgenson, D.W., Lau, L.J., Stoker, T.M. (1982). "The transcendental logarithmic model of aggregate consumer behavior". In: Basmann, R.L., Rhodes, G. (Eds.), Advances in Econometrics, vol. 1. JAI Press, Greenwich, pp. 97–238.

Jorgenson, D.W., Slesnick, D.T. (2005). "Consumer and labor supply". Manuscript. January.

Krusell, P., Smith, A.A. (1998). "Income and wealth heterogeneity in the macroeconomy". Journal of Political Economy 106 (5), 867–896.

Krusell, P., Smith, A. Jr. (2007). "Quantitative macroeconomic models with heterogeneous agents". In: Blundell, R.W., Newey, W.K., Persson, T. (Eds.), Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Volume I. Cambridge University Press, 2006. Cambridge Collections Online. Cambridge University Press. July.

Lau, L.J. (1977). "Existence conditions for aggregate demand functions". Institute for Mathematical Studies in the Social Sciences, Stanford University, Technical Report No. 248.

Lau, L.J. (1982). "A note on the fundamental theorem of exact aggregation". Economic Letters 9, 119–126.
Lewbel, A. (1989b). "Exact aggregation and a representative consumer". Quarterly Journal of Economics 104, 622–633.
Lewbel, A. (1990). "Income distribution movements and aggregate money illusions". Journal of Econometrics 43, 35–42.
Lewbel, A. (1991). "The rank of demand systems: Theory and nonparametric estimation". Econometrica 59, 711–730.
Lewbel, A. (1993). "Distributional movements, macroeconomic regularities and the representative consumer". Research in Economics 47, 189–199.
Lewbel, A. (1994). "Aggregation and simple dynamics". American Economic Review 84, 905–918.
Mace, B.J. (1991). "Full insurance in the presence of aggregate uncertainty". Journal of Political Economy 99, 928–956.
MaCurdy, T.E. (1982). "The use of time series processes to model the error structure of earnings in a longitudinal data analysis". Journal of Econometrics 18, 83–114 (January).
MaCurdy, T.E. (1987). "A framework for relating microeconomic and macroeconomic evidence on intertemporal substitution". In: Bewley, T.F. (Ed.), Advances in Econometrics, Fifth World Congress, vol. II, pp. 149–176.
McFadden, D., Reid, F. (1975). "Aggregate travel demand forecasting from disaggregated behavioral models". Transportation Research Record (534), 24–37.
Meghir, C., Pistaferri, L. (2004). "Income variance dynamics and heterogeneity". Econometrica 72 (1), 1–32.
Meghir, C., Weber, G. (1996). "Intertemporal nonseparability or borrowing restrictions? A disaggregate analysis using a US consumption panel". Econometrica 64 (5), 1151–1181.
Muellbauer, J. (1975). "Aggregation, income distribution and consumer demand". Review of Economic Studies 42, 525–543.
Pischke, J.-S. (1995). "Individual income, incomplete information and aggregate consumption". Econometrica 63, 805–840.
Pollak, R.A., Wales, T.J. (1981). "Demographic variables in demand analysis". Econometrica 49, 1533–1551.
Ray, R. (1983). "Measuring the cost of children: An alternative approach". Journal of Public Economics 22, 89–102.
Rogerson, R., Wallenius, J. (2007). "Micro and macro elasticities in a life cycle model with taxes". National Bureau of Economic Research, Working Paper 13017, Cambridge, MA, April.
Roy, A.D. (1951). "Some thoughts on the distribution of earnings". Oxford Economic Papers 3, 135–146.
Schafer, W., Sonnenschein, H. (1982). "Market demand and excess demand functions". In: Arrow, K.J., Intriligator, M.D. (Eds.), Handbook of Mathematical Economics, vol. II. North-Holland, New York.
Skinner, J. (1988). "Risky income. Life cycle consumption and precautionary savings". Journal of Monetary Economics 22, 237–255.
Solon, G., Barksy, R., Parker, J. (1994). "Measuring the cyclicality of real wages: How important is composition bias?". Quarterly Journal of Economics 109 (1), 1–26.
Sonnenschein, H. (1972). "Market excess demand functions". Econometrica 40, 549–563.
Stoker, T.M. (1984a). "Completeness, distribution restrictions and the form of aggregate functions". Econometrica 52, 887–907.
Stoker, T.M. (1984b). "Exact aggregation and generalized Slutsky conditions". Journal of Economic Theory 33, 368–377.
Stoker, T.M. (1986c). "Simple tests of distributional effects on macroeconomic equations". Journal of Political Economy 94, 763–795.
Stoker, T.M. (1993). "Empirical approaches to the problem of aggregation over individuals". Journal of Economic Literature 31, 1827–1874.
Townsend, R.M. (1994). "Risk and insurance in village India". Econometrica 62, 539–592.
Zeldes, S.P. (1989a). "Consumption and liquidity constraints: An empirical investigation". Journal of Political Economy 97 (2), 305–346.
Zeldes, S.P. (1989b). "Optimal consumption with stochastic income". Quarterly Journal of Economics 104, 275–298.

Further reading

Aigner, D.J., Goldfeld, S.M. (1974). "Estimation and prediction from aggregate data when aggregates are measured more accurately than their components". Econometrica 42, 113–134.

Allenby, G.M., Rossi, P.E. (1991). "There is no aggregation bias: Why macro logit models work". Journal of Economic and Business Statistics 9, 1–14.

Anderson, S.P., de Palma, A., Thisse, J.-F. (1989). "Demand for differentiated products, discrete choice models and the characteristics approach". Review of Economic Studies 56, 21–36.

Attanasio, O.P., Goldberg, P.K. (1997). "On the relevance of borrowing restrictions. Evidence from car loans in the US". NBER WP No. 7694.

Barker, T.S., Pesaran, M.H. (1990). Disaggregation in Econometric Modeling. Routledge, London.

Barnett, W.A. (1979). "Theoretical foundations of the Rotterdam model". Review of Economic Studies 46, 109–130.

Bean, C.R. (1986). "The estimation of 'surprise' models and the 'surprise' consumption function". Review of Economic Studies 53, 497–516.

Beaudry, P., Green, D. (1996). "Cohort patterns in Canadian earnings and the skill-based technical change hypothesis". Working Paper. Department of Economics, University of British Columbia.

Becker, G.S. (1962). "Irrational behavior and economic theory". Journal of Political Economy 70, 1–13.

Berndt, E.R., Darrough, M.N., Diewert, W.E. (1977). "Flexible functional forms and expenditure distributions: An application to Canadian consumer demand functions". International Economic Review 18, 651–675.

Bewley, T. (1977). "The permanent income hypothesis: A theoretical formulation". Journal of Economic Theory 16, 252–292.

Bils, M.J. (1985). "Real wages over the business cycle: Evidence from panel data". Journal of Political Economy 93, 666–689.

Blackorby, C., Boyce, R., Russell, R.R. (1978). "Estimation of demand systems generated by the Gorman polar form: A generalization of the S-branch utility tree". Econometrica 46, 345–364.

Blinder, A.S. (1975). "Distribution effects and the aggregate consumption function". Journal of Political Economy 83, 447–475.

Blundell, R.W., Ham, J., Meghir, C. (1987). "Unemployment and female labour supply". Economic Journal 97, 44–64.

Blundell, R.W., MaCurdy, T. (1999). "Labor supply: A survey". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. III(a). North-Holland, Amsterdam.

Blundell, R.W., Preston, I. (1998). "Consumption inequality and income uncertainty". Quarterly Journal of Economics 113 (2), 603–640 (May).

Blundell, R.W., Robin, J.-M. (1999). "Estimation in large and disaggregated demand systems: An estimator for conditionally linear systems". Journal of Applied Econometrics 14 (3), 209–232 (May–June).

Browning, M.J. (1993). "Estimating micro parameters from macro data alone: Some pessimistic evidence". Research in Economics 47, 253–268.

Browning, M.J., Lusardi, A. (1996). "Household saving: Micro theories and micro facts". Journal of Economic Literature 34 (4), 1797–1855.

Browning, M.J., Meghir, C. (1991). "The effects of male and female labour supply on commodity demands". Econometrica 59, 925–951.

Buse, A. (1992). "Aggregation, distribution and dynamics in the linear and quadratic expenditure systems". Review of Economics and Statistics 74, 45–53.

Caballero, R.J., Engel, E.M.R.A. (1991). "Dynamic (S,s) economies". Econometrica 59, 1659–1686.

Caballero, R.J., Engel, E.M.R.A. (1992). Microeconomic Adjustment Hazards and Aggregate Demands. MIT Department of Economics, Cambridge, MA.

Cameron, A.C. (1990). "Aggregation in discrete choice models: An illustration of nonlinear aggregation". In: Barker, T.S., Pesaran, M.H. (Eds.), Disaggregation in Econometric Modeling. Routledge, London.

Caplin, A.S., Nalebuff, B. (1991). "Aggregation and imperfect competition: On the existence of equilibrium". Econometrica 59, 1–24.

Caplin, A.S., Spulber, D. (1987). "Menu costs and the neutrality of money". Quarterly Journal of Economics 102, 703–726.
Chetty, V.K., Heckman, J.J. (1986). "A dynamic model of aggregate output supply, factor demand and entry and exit for a competitive industry with heterogeneous plants". Journal of Econometrics 33, 237–262.
Cogan, J.F. (1981). "Fixed costs and labor supply". Econometrica 49, 945–964.
Cowing, T.G., McFadden, D.L. (1984). Microeconomic Modeling and Policy Analysis. Academic Press, New York.
de Wolff, P. (1941). "Income elasticity of demand, a micro-economic and a macro-economic interpretation". Economic Journal 51, 140–145.
Deaton, A.S. (1985). "Panel data from time series of cross sections". Journal of Econometrics 30, 109–126.
Deaton, A.S. (1991). "Savings and liquidity constraints". Econometrica 59, 1221–1248.
Deaton, A.S. (1993). Understanding Consumption. Oxford University Press, Oxford.
Debreu, G. (1974). "Excess demand functions". Journal of Mathematical Economics 1, 15–23.
Dickens, R. (1996). "The evolution of individual male earnings in Great Britain, 1975–1994". CEP Discussion Paper No. 306, November.
Dumas, B. (1989). "Two-person dynamic equilibrium in the capital market". The Review of Financial Studies 2, 159–188.
Epstein, L.G., Zin, S.E. (1989). "Substitution, risk aversion and the temporal behaviour of consumption and asset returns: A theoretical framework". Econometrica 57 (4), 937–969.
Epstein, L.G., Zin, S.E. (1991). "Substitution risk aversion and the temporal behaviour of consumption and asset returns: An empirical analysis". Journal of Political Economy 99 (2), 263–286.
Fair, R.C., Dominguez, K.M. (1991). "Effects of the changing US age distribution on macroeconomic equations". American Economic Review 81, 1276–1294.
Feenberg, D.R., Rosen, H.S. (1983). "Alternative tax treatment of the family: Simulation methodology and results". In: Feldstein, M.S. (Ed.), Behavioral Simulation Methods in Tax Policy Analysis. University of Chicago Press, Chicago. Chapter 1.
Freixas, X., Mas-Colell, A. (1987). "Engel curves leading to the weak axiom in the aggregate". Econometrica 55, 515–532.
Gosling, A., Machin, S., Meghir, C. (2000). "The changing distribution of male wages in the UK, 1966–1993". Review of Economic Studies 67 (4), 635–666.
Grunfeld, Y., Griliches, Z. (1960). "Is aggregation necessarily bad?". Review of Economics and Statistics 42, 1–13.
Hall, R.E. (1988). "Intertemporal substitution in consumption". Journal of Political Economy 96 (2), 339–357.
Hansen, G.D. (1985). "Indivisible labor and the business cycle". Journal of Monetary Economics 16, 309–327.
Hansen, L.P., Singleton, K.J. (1982). "Generalized instrumental variables estimation of nonlinear rational expectations models". Econometrica 50 (5), 1269–1286.
Hansen, L.P., Singleton, K.J. (1983). "Stochastic consumption, risk aversion and the temporal behavior of asset returns". Journal of Political Economy 91, 249–265.
Hausman, J.A., Newey, W.K., Ichimura, H., Powell, J.L. (1991). "Identification and estimation of polynomial errors in variables models". Journal of Econometrics 50, 273–296.
Hayashi, F. (1985). "The effects of liquidity constraints on consumption: A cross section analysis". Quarterly Journal of Economics 100, 183–206.
Heckman, J.J. (1974). "Effects of child-care programs on women's work effort". Journal of Political Economy 82 (2), S136–S163.
Heckman, J.J. (1979). "Sample selection bias as a specification error". Econometrica 47, 153–161.
Heckman, J.J. (1990). "Varieties of selection bias". American Economic Review 80 (2), 313–318.
Heckman, J.J., Chetty, V.K. (1986). "A dynamic model of aggregate output, factor demand and entry and exit for a competitive industry with heterogeneous plants". Journal of Econometrics 33, 237–262.
Heckman, J.J., Walker, J.R. (1989). "Forecasting aggregate period-specific birth rates: The time series properties of a microdynamic neoclassical model of fertility". Journal of the American Statistical Association 84, 958–965.

Heckman, J.J., Walker, J.R. (1990). "The relationship between wages and income and the timing and spacing of births: Evidence from Swedish longitudinal data". Econometrica 58, 1411–1442.
Heineke, J.M., Shefrin, H.M. (1990). "Aggregation and identification in consumer demand systems". Journal of Econometrics 44, 377–390.
Hicks, J.R. (1956). A Revision of Demand Theory. Oxford University Press, London.
Hildenbrand, W. (1981). "Short run production functions based on microdata". Econometrica 49, 1095–1126.
Hildenbrand, W. (1983). "On the law of demand". Econometrica 51, 997–1019.
Hildenbrand, W. (1992). "Market demand, theory and empirical evidence". Universitat Bonn Discussion Paper No. A-359.
Houthakker, H.S. (1955). "The Pareto distribution and the Cobb–Douglas production function in activity analysis". Review of Economic Studies 23, 27–31.
Johansen, L. (1972). Production Functions. North-Holland, Amsterdam.
Joint Committee on Taxation (1992). Discussion of Revenue Estimation and Process, JCS 14-92, August 13, 1992. US Government Printing Office, Washington, DC.
Jorgenson, D.W., Slesnick, D.T. (1984). "Aggregate consumer behavior and the measurement of inequality". Review of Economic Studies 51, 369–392.
Jorgenson, D.W., Slesnick, D.T., Stoker, T.M. (1988). "Two stage budgeting and exact aggregation". Journal of Business and Economic Statistics 6, 313–325.
Jorgenson, D.W., Stoker, T.M. (1986). "Nonlinear three stage least squares pooling of cross section and average time series data". In: Moroney, J.R. (Ed.), Advances in Statistical Analysis and Statistical Computing, vol. 1. JAI Press, Greenwich, pp. 87–116.
Katz, L.F., Murphy, K.M. (1992). "Changes in relative wages, 1963–1987: Supply and demand factors". Quarterly Journal of Economics 107 (1), 35–78.
Kelejian, H.J. (1980). "Aggregation and disaggregation of nonlinear equations". In: Kmenta, J., Ramsey, J.B. (Eds.), Evaluation of Econometric Models. Academic Press, New York.
Kirman, A.P. (1992). "Whom or what does the representative individual represent?". Journal of Economic Perspectives 6, 117–136.
Lam, P.-S. (1992). "Permanent income, liquidity, and adjustments of automobile stocks: Evidence from panel data". Quarterly Journal of Economics 106, 203–230.
Lee, K., Pesaran, M.H., Pierse, R. (1990). "Testing for aggregation bias in linear models". Economic Journal 100, 137–150.
Lee, L.-F., Porter, R.H. (1984). "Switching regression models with imperfect sample separation: With an application to cartel stability". Econometrica 52, 391–418.
Leser, C.E.V. (1963). "Forms of Engel functions". Econometrica 31, 694–703.
Lewbel, A. (1989a). "A demand system rank theorem". Econometrica 57, 701–706.
Lewbel, A. (1992). "Aggregation with log-linear models". Review of Economic Studies 59, 635–642.
Lippi, M. (1988). "On the dynamic shape of aggregated error correction models". Journal of Economic Dynamics and Control 12, 561–585.
Lusardi, A. (1996). "Permanent income, current income and consumption: Evidence from two panel data sets". Journal of Business and Economic Statistics 14, 81–89.
Mankiw, N.G., Weil, D.N. (1989). "The baby boom, the baby bust, and the housing market". Regional Science and Urban Economics 19, 235–258.
Marshak, J. (1939). "Personal and collective budget functions". Review of Economic Statistics 21, 161–170.
Meghir, C., Whitehouse, E. (1996). "The evolution of wages in the UK: Evidence from micro data". Journal of Labor Economics 14 (1), 1–25.
Modigliani, F. (1970). "The life cycle hypothesis of saving and intercountry differences in the saving ratio". In: Eltis, W., Scott, M., Wolfe, I. (Eds.), Induction, Growth and Trade. Clarendon Press, Oxford.
Moffitt, R. (1993). "Identification and estimation of dynamic models with a time series of repeated cross sections". Journal of Econometrics 59 (1–2), 99–123.
Muellbauer, J. (1976). "Community preferences and the representative consumer". Econometrica 44, 979–1000.

Muellbauer, J. (1981). "Linear aggregation in neoclassical labor supply". Review of Economic Studies 48, 21–36.

Pesaran, M.H., Pierce, R.G., Kumar, M.S. (1989). "Econometric analysis of aggregation in the context of linear prediction models". Econometrica 57, 861–888.

Pollak, R.A. (1985). "A transactions cost approach to families and households". Journal of Economic Literature 23, 581–608.

Powell, J.L. (1994). "Estimation of semi-parametric models". In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. Elsevier, Amsterdam.

Powell, J.L., Stoker, T.M. (1985). "Estimation of complete aggregation structures". Journal of Econometrics 30, 317–344.

Prescott, E.C. (1986). "Theory ahead of business cycle measurement". Quarterly Review 10, 9–22. Federal Reserve Bank of Minneapolis.

Rubinstein, M. (1974). "An aggregation theorem for securities markets". Journal of Financial Economics 1, 224–244.

Runkle, D.E. (1991). "Liquidity constraints and the permanent income hypothesis: Evidence from panel data". Journal of Monetary Economics 27, 73–98.

Samuelson, P.A. (1956). "Social indifference curves". Quarterly Journal of Economics 73, 1–22.

Sato, K. (1975). Production Functions and Aggregation. North-Holland, Amsterdam.

Scheinkman, J.A., Weiss, L. (1986). "Borrowing constraints and aggregate economic activity". Econometrica 54 (1), 23–45.

Stoker T.M. (1978). "The pooling of cross section and average time series data". Draft. Harvard University.

Stoker, T.M. (1982). "The use of cross section data to characterize macro functions". Journal of the American Statistical Association 77, 369–380.

Stoker, T.M. (1985). "Aggregation, structural change and cross section estimation". Journal of the American Statistical Association 80, 720–729.

Stoker, T.M. (1986a). "Aggregation, efficiency and cross section regression". Econometrica 54, 177–188.

Stoker, T.M. (1986b). "The distributional welfare impacts of rising prices in the United States". American Economic Review 76, 335–349.

Stoker, T.M. (1986d). "Consistent estimation of scaled coefficients". Econometrica 54, 1461–1482.

Stoker, T.M. (1992). Lectures on Semiparametric Econometrics. CORE Lecture Series. CORE Foundation, Louvain-la-Neuve.

Theil, H. (1954). Linear Aggregation of Economic Relations. North-Holland, Amsterdam.

Theil, H. (1975). Theory and Measurement of Consumer Demand, Vol. 1. North-Holland, Amsterdam.

Van Daal, J., Merkies, A.H.Q.M. (1984). Aggregation in Economic Research. D. Reidel, Dordrecht.

Weber, G. (1993). "Earnings related borrowing restrictions: Empirical evidence from a pseudo panel for the UK". Annales d'Economie et de Statistique 29, 157–173.

Working, H. (1943). "Statistical laws of family expenditure". Journal of the American Statistical Association 38, 43–56.

Zellner, A. (1969). "On the aggregation problem, a new approach to a troublesome problem". In: Estimation and Risk Programming, Essays in Honor of Gerhard Tintner. Springer, Berlin.

Chapter 69

LABOR SUPPLY MODELS: UNOBSERVED HETEROGENEITY, NONPARTICIPATION AND DYNAMICS*

RICHARD BLUNDELL
University College London, UK
Institute for Fiscal Studies, London, UK

THOMAS MACURDY
Department of Economics, Stanford University, USA
Hoover Institute, Stanford University, USA

COSTAS MEGHIR
University College London, UK
Institute for Fiscal Studies, London, UK

Contents

* We would like to thank Martin Browning, the editors and participants at the London Handbook meeting for helpful comments. This study is part of the research program of the ESRC Centre for the Microeconomic Analysis of Fiscal Policy at the IFS.

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06069-2

Abstract

This chapter is concerned with the identification and estimation of models of labor supply. The focus is on the key issues that arise from unobserved heterogeneity, nonparticipation and dynamics. We examine the simple "static" labor supply model with proportional taxes and highlight the problems surrounding nonparticipation and missing wages. The difference-in-differences approach to estimation and identification is developed within the context of the labor supply model. We also consider the impact of incorporating nonlinear taxation and welfare program participation. Family labor supply is looked at from both the unitary and collective perspectives. Finally we consider intertemporal models focusing on the difficulties that arise with participation and heterogeneity.

Keywords

labor supply, consumption, taxation, microeconometrics

JEL classification: D1, D9, J2

1. Introduction

This chapter is concerned with the identification and estimation of labor supply models. The specification and estimation of such models has already been the subject of numerous studies and surveys.[1] So why this one? The overall objective of this chapter is to consider models that allow policy evaluation and simulation allowing for individual heterogeneity. Evaluation concerns the assessment of reforms that have taken place. Policy simulation concerns the assessment of proposed reforms. For the most part it is the latter that has been the central concern of empirical researchers. That is to construct a model that can reliably be used for the assessment of proposed reforms. Since many policy proposals involve the reform of highly nonlinear budget constraints and impact decisions that are discrete and cover the whole life-cycle, we argue that a fully specified dynamic structural model is the ideal. In particular, it is of central importance to consider how labor supply and saving decisions interact and how policy affects labor supply decisions within a period as well as intertemporally. However, this ideal has a number of practical and theoretical difficulties. In certain situations, the evaluation of existing reforms can be analyzed using much simpler and potentially more robust techniques.

To best convey the set of issues surrounding estimation of labor supply models we start with the simplest static framework and build up to the more complete dynamic models, adding important elements such as nonlinear budget sets on the way. Thus, the layout of the chapter is as follows. Section 2 presents an assessment of the estimation issues underlying the simple 'static' labor supply model with proportional taxes and highlights the problems surrounding nonparticipation and missing wages. In Section 3 we consider the natural experiment and difference-in-differences approaches to estimation and evaluation of reforms, laying out the identifying assumptions underlying interpretation of the results. We consider estimation of a simple discrete policy response parameter as well as the estimation of income and substitution effects. In Section 4 we examine the impact of incorporating nonlinear taxation and welfare program participation. Section 5 considers some of the specific issues that relate to family labor supply, including the development of the collective approach and welfare program participation as previously articulated. Section 6 discusses intertemporal labor supply models. This section reviews the various approaches taken to dynamic modeling and examines the difficulties that arise with participation and heterogeneity. Section 7 concludes the chapter.

2. Estimation and identification with participation with proportional taxes

We begin by considering the simple static model of hours and consumption choices. We leave the discussion of nonlinear budget sets to Section 4.

[1] For overall evaluations and surveys see Killingsworth (1983), MaCurdy (1985), Killingsworth and Heckman (1986), Heckman (1993), Mroz (1987), Hausman (1985a, 1985b), Pencavel (1986), Blundell and MaCurdy (1999) to mention a few.

2.1. Static specifications

2.1.1. The allocation of hours and consumption

Utility is defined over hours of work h and consumption c, both of which are restricted to be nonnegative and h is restricted to be below a maximal amount of an available time endowment. Formally, this discussion is easily extended to the case of family labor supply decisions where h is a vector of household labor supplies. However, there are many specific issues relating to joint participation decisions and to the allocation of resources within the family that are central to any study of family labor supply; we leave our discussion of family labor supply models to Section 5. Equally, consumption decisions can be disaggregated. This disaggregation is central to the analysis of nonseparability of goods and leisure. We turn to this below.

If we let y represent the total unearned income available for consumption, and w the real wage rate, then the optimal choices for c and h are given by the solution to

$$\max_{c,h}\{U(c,h) \mid c - wh = y;\ c \geqslant 0;\ h \geqslant 0\} \tag{2.1}$$

where $U(c,h)$ is a quasiconcave utility index defined increasing in c and $-h$. The resulting labor supply has the form

$$h = h(w, y). \tag{2.2}$$

In the static model y is taken to be income from other sources. However it turns out that the precise definition of y is crucial: If y is measured as the difference between total consumption expenditure and earnings, $c - wh = y$, it is consistent both with intertemporal two-stage budgeting both in the absence of liquidity constraints and with the presence of liquidity constraints that are unrelated to labor supply. This is discussed in a subsection below.

The indirect utility representation of preferences is given by

$$V(w, y) \equiv U\big(wh(w, y) + y, h(w, y)\big) \tag{2.3}$$

which is linear homogeneous, quasiconcave in prices $\mathbf{p}$, w and y, decreasing in $\mathbf{p}$ and w and increasing in y. The various representations of preferences (direct or indirect utility) detailed below are going to be particularly useful in specifying empirical models and defining the likelihood function.

2.1.2. Two-stage budgeting specifications and within-period allocations

Labor supply and consumption models are frequently analyzed in a two-good framework. Such modeling is less restrictive than it sounds because under Gorman's (1959, 1968) two-stage budgeting, this labor supply model can be seen as the top stage where "full income" is shared between consumption and leisure and then the consumption

budget is split among goods. However, for such an interpretation with all goods being represented by one or two price indices, we require some conditions on preferences.

Suppose utility is defined over hours of work h and a vector of goods $\mathbf{q}$. Assume the individual has a within-period utility function of the form

$$\upsilon_t = v(c_t, h_t, \mathbf{p}_t) = \max_{\mathbf{q},h} \{u(\mathbf{q}_t, h_t) \mid \mathbf{p}_t'\mathbf{q}_t = c_t\} \tag{2.4}$$

where $\mathbf{p}_t$ is a vector of prices corresponding to the disaggregated commodity vector $\mathbf{q}_t$. The function υ_t is a conditional indirect utility function which is increasing in total consumption expenditure c_t, decreasing and concave in prices and decreasing in hours of work h_t.

We say that $\mathbf{q}_t$ is weakly separable from h_t if the marginal rate of substitution between goods $\mathbf{q}_t$ does not depend on h_t. In this case the utility function can be written as $u(u_1(\mathbf{q}_t), h_t)$ where u_1 is a sub-utility function. If in addition the marginal utilities of $\mathbf{q}_t$ and h_t do not depend on each other then we say that the utility function is additively separable, in which case the utility function can be written as $u_1(\mathbf{q}_t) + u_2(h_t)$. Blackorby, Primont and Russell (1978) have a detailed analysis of the concepts of separability and Deaton (1974) analyzes the empirical implications of the additive separability assumption.

Gorman (1959) has shown that if a set of goods x_1 is separable from goods x_2 then it is possible to express the demands for goods x_1 simply as a function of the total expenditure allocated to this group (x_1) and the prices of these goods alone (say p_1). In addition, if preferences can be expressed in the generalized Gorman polar form, then it is possible to express the overall expenditure allocations to each group as a function of the price indices for each group. This theorem can justify considering the allocation of total expenditure to overall consumption and leisure separately from the problem of how expenditure is allocated to goods. However, it has to be borne in mind that the justification which allows us to write labor supply as a function of the real wage alone (rather than of all relative prices) does imply restrictions on preferences.

These results offer a justification of the static model within an intertemporal context since the concept of separability can extend both over goods and over time.[2] Typically we impose additive separability over time in which case the marginal utility of consumption or hours of work in one period is unaffected by consumption and hours in any other time period. Additive intertemporal separability has the implication that we can use two-stage budgeting to characterize consumption choices: given the level of consumption and separability, the within-period demands for goods $\mathbf{q}_t$ only depend on the prices of those goods and on wages (if the goods are not separable from hours). The indirect utility function defined by (2.4) then becomes the criterion function for allocating consumption (and hours) over the life-cycle.[3]

[2] See Gorman (1959), MaCurdy (1983), Altonji (1986), Blundell and Walker (1986) and Arellano and Meghir (1992).

[3] Utility (2.4) implicitly assumes separability over time thus ruling out habits and/or adjustment costs [see Hotz, Kydland and Sedlacek (1988) and Meghir and Weber (1996)].

It is well known that taking a monotonic transformation of the utility function does not change the observed within-period allocations. In an intertemporal context this issue acquires a special importance: taking a monotonic transformation does not alter the way that consumption and hours are allocated within period, under intertemporal separability. However, it does potentially change the marginal rate of substitution between periods. Hence, as we will discuss further below, estimating intertemporal preferences generally requires intertemporal data.

Noting that modeling the monotonic transformation is modeling intertemporal preferences, we use the slightly more elaborate notation

$$v_t = \psi\big[U(c_t, h_t|z_{1t}), z_{2t}\big] \tag{2.5}$$

where $\psi[\cdot]$ is a monotonic function of its first argument U and where z_1 and z_2 are variables (observed or otherwise) that affect preferences over consumption and hours of work. In particular, z_{2t} affects intertemporal allocations but not within-period ones (unless it contains common elements with z_{1t}). Our focus in this section is on within-period allocations. The discussion here should make it clear that one can work with the utility function (2.1) to represent within-period allocations of consumption and hours of work consistent with life-cycle choices.

2.1.3. *Empirical labor supply specifications*

Preferences can be represented by direct utility functions, indirect utility functions or by the labor supply equation itself. In each case the function has to satisfy some basic properties to be consistent with theory. Here we briefly review some standard specifications of the static labor supply model (2.2) and relate them to their indirect utility function. Such specifications are usually chosen for ease of estimation and here we simply consider the specifications and their underlying model of preferences. With unobserved heterogeneity and nonparticipation it is useful, if not essential, to have some relatively simple parametric specification in mind.

The *linear labor supply* model

$$h = \alpha + \beta w + \gamma y \tag{2.6}$$

has indirect utility

$$V(w, y) = e^{\gamma w}\left(y + \frac{\beta}{\gamma}w - \frac{\beta}{\gamma^2} + \frac{\alpha}{\gamma}\right) \quad \text{with } \gamma \leqslant 0 \text{ and } \beta \geqslant 0. \tag{2.7}$$

Although popular [see Hausman (1981, 1985a, 1985b), for example], it is arguable that this linear specification allows too little curvature with wages.

Alternative *semilog specifications* and their generalizations are also popular in empirical work. For example, the semilog specification

$$h = \alpha + \beta \ln w + \gamma y \tag{2.8}$$

with indirect utility

$$V(w, y) = \frac{e^{\gamma w}}{\gamma}\left(\gamma y + \alpha\frac{\beta}{\gamma} + \beta \ln w\right) - \frac{\beta}{\gamma}\int_{\gamma y}\frac{e^{\gamma y}}{\gamma y}\, d(\gamma y)$$
$$\text{with } \gamma \leqslant 0 \text{ and } \beta \geqslant 0. \tag{2.9}$$

Moreover, the linearity of (2.8) in α and $\ln w$ makes it particularly amenable to an empirical analysis with unobserved heterogeneity, endogenous wages and nonparticipation as discussed below. Consequently, this specification is used extensively in our discussion of estimation that follows.

Neither (2.6) nor (2.8) allows backward-bending labor supply behavior although it is easy to generalize (2.8) by including a quadratic term in $\ln w$. Note that imposing integrability conditions at zero hours for either (2.6) or (2.8) implies positive wage and negative income parameters. A simple specification that does allow backward-bending behavior, while retaining a three-parameter linear-in-variables form, is that used in Blundell, Duncan and Meghir (1992):

$$h = \alpha + \beta \ln w + \gamma\frac{y}{w} \tag{2.10}$$

with indirect utility

$$V(w, y) = \frac{w^{\beta+1}}{\beta+1}\left(\frac{y}{w}(1+\gamma)^2 + \beta \ln w + \alpha - \frac{\beta}{1+\gamma}\right) \quad \text{with } \gamma \leqslant 0 \text{ and } \beta \geqslant 0. \tag{2.11}$$

This form has similar properties to the MRS specification of Heckman (1974c).

Generalizations of the *Linear Expenditure System or Stone–Geary preferences* are also attractive from certain points of view. For example suppose the indirect utility function for individual i in period t takes the form

$$V_{it} = \left[\frac{wH + y - a(w)}{b(w)}\right] \tag{2.12}$$

where H is the maximum amount of hours available to be allocated between hours and leisure. This is the quasi-homothetic "Gorman polar form". The linear expenditure system belongs to this class. However, there is no need to impose additive separability between consumption and hours of work as would be the case under Stone–Geary/LES preferences. Indeed, such separability assumptions severely constrain the time path of consumption and hours of work and can lead to the impression that the life-cycle model is unable to explain a number of observed phenomena, see Heckman (1974b). In particular we may specify

$$a(w) = a_0 + a_1 w + 2a_2 w^{\frac{1}{2}} \tag{2.13}$$

and

$$b(w) = w^{\beta} \tag{2.14}$$

which is a *Generalized Leontief* model. Preferences are additive and reduce to LES if $a_2 = 0$.

The implied labor supply function using (2.12)–(2.14) can be derived using Roy's identity and takes the form

$$h_{it} = (H - a_1) - a_2 w^{-\frac{1}{2}} - \frac{\beta}{w}\left(M - a_0 + a_1 w + 2a_2 w^{\frac{1}{2}}\right) \tag{2.15}$$

where $M = wH + y$. Unobserved heterogeneity can also easily be allowed for, as well as measurement error in hours of work (but not in hourly wages) and/or consumption. For example, we can allow a_1 to be heterogeneous across individuals and time, i.e. $a_1 = \bar{a}_1 + \varepsilon$. Under the simplifying assumption that a_1 is the only source of heterogeneity the error term in the earnings equation now becomes $\nu = -\varepsilon(1 + \beta)$.

2.2. *Estimation of the static labor supply model*

The main estimation issue, ignoring problems related to participation and nonlinear taxation (discussed below), is the endogeneity of wages w and unearned income y. Wages may well be endogenous because unobservables affecting preferences for work may well be correlated with unobservables affecting productivity and hence wages. Unearned income may be endogenous for a number of reasons: If y represents asset income, then individuals who work harder (because of unobserved preferences for work) are also likely to have accumulated more assets.[4]

Take as a simple example the semilog model of labor supply as above, i.e.

$$h_i = \alpha' x_i + \beta \ln w_i + \gamma y_i + u_i. \tag{2.16}$$

The subscript i denotes an individual. The variables x denote observables which determine preferences. We avoid using the log of y because it is conceivable that it is zero and, in some cases, even negative. We add to this system a wage equation

$$\ln w_i = \delta_1' x_i + \delta_2' z_i + v_i$$

and a reduced form equation for unearned income

$$y_i = \zeta_1' x_i + \zeta_2' z_i + \varepsilon_i.$$

Identification requires that the dimension of the variables excluded from the labor supply equation, z_i, is at least two. It also requires that the matrix $[\delta_2' \zeta_2']$ has rank 2. In this linear framework, estimation is straightforward – two-stage least squares is the obvious choice. However, we will see below that it is convenient to estimate the three reduced forms first and then impose the parametric restrictions to recover the structural

[4] If μ also represents income from spouses, positive assortative mating will imply that hard-working individuals will tend to marry. Hence unobserved preferences for work will correlate with spousal income reflected in μ.

coefficients using minimum distance. The reduced form labor supply model is

$$h_i = (\alpha + \beta\delta_1 + \gamma\zeta_1)'x_i + (\beta\delta_2 + \gamma\zeta_2)'z_i + u_i.$$

Given estimates of all the reduced form coefficients the restrictions can then be imposed using minimum distance. Thus let

$$\alpha_1 = (\alpha + \beta\delta_1 + \gamma\zeta_1), \qquad \alpha_2 = (\beta\delta_2 + \gamma\zeta_2), \qquad \alpha_3 = \left[\delta_1'\delta_2'\zeta_1'\zeta_2'\right]',$$

and let Ω represent the covariance matrix of the OLS estimator of the three-equation reduced form system. Finally let $\alpha(\theta) = [\alpha_1\alpha_2\alpha_3]'$ where θ represents the set of parameters in the labor supply model, the wage equation and the unearned income equation. Then the optimal minimum distance estimator is

$$\hat{\theta} = \arg\min_{\theta}\left\{\left(\hat{\alpha} - \alpha(\theta)\right)'\Omega^{-1}\left(\hat{\alpha} - \alpha(\theta)\right)\right\}.$$

The resulting estimator is efficient, to the extent that the first-step estimator is efficient.

When the labor supply model is nonlinear this straightforward procedure is no longer available. In this case an alternative approach is maximum likelihood or semiparametric instrumental variables. Maximum likelihood will be discussed below in the context of the labor supply model with corner solutions and nonlinear taxation. Hence we avoid duplication by deferring discussion until then.

In the absence of censoring, one can use nonparametric instrumental variables as in Newey and Powell (2003) and Darolles, Florens and Renault (2000). Consider the case where the labor supply is an unknown function of w and y

$$h_i = h(w_i, y_i) + u_i.$$

The object is to estimate the function h. Suppose we have a set of instruments z (at least two if we are to treat both the wage and other income as endogenous). We assume that the error in the labor supply function satisfies the rank condition, $E(u_i|z_i) = 0$. In addition one needs a strong identification assumption ensuring that *any* function of w, y can be explained by the instruments z. Under these conditions solving the moment condition

$$E\left(h_i - h(w_i, y_i)|z_i\right) = 0$$

for the function $h(w_i, y_i)$ provides a nonparametric estimator.

In the context of censoring due to nonparticipation a control function approach turns out to be more useful. However, it is important to note that the assumptions underlying the control function are different from those underlying the IV approach above, unless the instruments are independent of the unobservables.[5] A form of the control function approach relies on the assumption that

$$E(u_i|z, x, w, y) = g(v_i, \varepsilon_i)$$

[5] Florens et al. (2007).

where v_i and ε_i are the error terms from the wage and unearned income equations respectively.[6] With unknown h, identification also requires measurable separability which ensures that the functions g and h vary independently and is the equivalent of the rank condition. In a parametric framework the requirements are less stringent since we are restricting attention to specific functional forms. One approach to estimation would be to take a series expansion of g. Alternatively we could use some kernel estimator. The procedure works under a generalized rank condition; however the important point to note is that even under nonlinearity we do not require explicit distributional assumptions, other than the restriction on the conditional expectation of u.[7] Nevertheless it should be noted that in practice it may be difficult to motivate the control function assumption, which contrasts with the orthogonality conditions above that are often derived from economic theory.

2.3. The censored regression model

Labor market participation raises two key questions for modeling labor supply. First, what market wage distribution should be used for nonparticipants? Second, are there features of the labor market that make labor supply behavior on the extensive margin (participation) fundamentally different from behavior on the intensive margin (hours of work)? These questions are not wholly unrelated since, without further restrictions on the distribution of offered wages among nonparticipants, it is difficult to separately identify a process for nonparticipation and for hours of work.

Among the most compelling reasons for separating these two margins is fixed costs of work – either monetary or time. We take up the issue of fixed costs in Section 2.5, and begin by working through a model without fixed costs. We consider first semiparametric estimation in a model with missing wages.

Suppose individual heterogeneity in tastes for work is represented by the random variable v. Observed hours of work (2.2) in the censored regression case can be represented by

$$h = \max\{f(w, y, x, v), 0\} \tag{2.17}$$

where $f(w, y, x, v)$ represents desired hours of work

$$f(w, y, x, v) \equiv h^* \tag{2.18}$$

and where y represents some measure of current period unearned income.

The censored labor supply model implies the reservation wage condition

$$h > 0 \quad \Leftrightarrow \quad w > w^*(y, x, v) \tag{2.19}$$

[6] For a more general case with unknown h see Newey, Powell and Vella (1999) or Florens et al. (2007) who derive conditions for identification.

[7] Two functions $g(e)$ and $h(v)$ are measurably separable iff whenever $g(e) - h(v) = 0$ a.s. implies $g(e)$ and $h(v)$ are constant functions.

where w^* is defined implicitly by

$$0 = f(w^*, y, x, v). \tag{2.20}$$

The existence and uniqueness of the reservation wage in this simple world is guaranteed by revealed preference arguments. Given the market wage w, (2.17) also defines a threshold condition on the unobservable heterogeneity term v given by

$$h > 0 \quad \Leftrightarrow \quad v \geqslant v^*(w, y, x) \quad \Rightarrow \quad \Pr(h > 0) = \int_{v \geqslant v^*} g(v)\, dv$$

where $g(v)$ is the density function for v.

To implement this censored regression specification we define the index I_i as an indicator variable that is unity if individual i participates[8] and zero otherwise. Observable hours of work then follow the rule

$$h_i = \begin{cases} h_i^* & \text{if } I_i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.21}$$

That is

$$\begin{aligned} I_i = 1 \quad &\Leftrightarrow \quad h_i^* > 0 \\ &= 1\{h_i^* > 0\}. \end{aligned} \tag{2.22}$$

This implies that participation in work follows a simple *corner-solution* framework and is equivalent to assuming there are no fixed costs.[9]

The log likelihood for an independently distributed random sample of n individuals in the censored model is given by

$$\ln L(\theta) = \sum_{i=1}^{n} \left(I_i \ln g(v; \theta) + (1 - I_i) \ln \int_{v \geqslant v^*} g(v; \theta)\, dv \right) \tag{2.23}$$

where θ are the unknown parameters of preferences and g is the distribution of v. In a linear specification with a normal iid assumption on v, this is equivalent to the Tobit censored regression specification.

The likelihood specification (2.23) makes two implicit assumptions on the wage distribution. First, that wages are observed for all individuals irrespective of their labor market status. Second, that wages are exogenous for labor supply. Neither of these is a priori reasonable.

[8] By participation we mean participation in paid work.

[9] By contrast, the fixed costs framework retains

$$I_i = 1 \quad \Rightarrow \quad h_i^* > 0$$

but not the reverse. As Cogan (1980) shows, fixed costs are equivalent to a positive reservation hours of work. We elaborate on this in Section 2.5 below.

2.4. Missing wages

Wages are not observed if $h = 0$. Suppose the model for wages can be written as

$$\ln w = \gamma_1 x + \gamma_2 q + \eta \tag{2.24}$$

where q are a set of variables that are exclusive to the determination of real wages and where η is an iid error term with distribution $g_w(\eta)$. The likelihood contribution for $h = 0$ becomes

$$\begin{aligned} &h > 0, \quad \ell_0 = g(v) g_w(\eta), \\ &h = 0, \quad \ell_0 = \int^{\infty} \int^{v^*} g(v) g_w(\eta)\, dv\, d\eta. \end{aligned} \tag{2.25}$$

By writing the joint distribution of v and η as a product of the two marginals we have implicitly maintained that wages are exogenous for labor supply. This implies that the density of wages can be estimated separately; in a labor supply model linear in log wages this further implies that we can simply impute wages for all nonworkers and estimate the model as if wages are observed (correcting the standard errors of course for generated regressor bias). However, if we wish to relax this assumption and permit w to be endogenous in the hours equation, the sample likelihood becomes

$$\ln L(\phi) = \sum_{i=1}^{n} \left(I_i \ln g_{hw}(v, \eta) + (1 - I_i) \ln \int^{\infty} \int^{v^*} g_{hw}(v, \eta)\, dv\, d\eta \right) \tag{2.26}$$

where $g_{hw}(v, \eta; \phi)$ is the joint distribution of v and η.

The resulting estimator simplifies enormously if we assume a parametric specification that permits an explicit reduced form for desired hours of work. A popular example of such a specification is the semilog labor supply model to which we now turn.

2.4.1. A semilog specification

Suppose we write the optimal labor supply choice for individual i as

$$h_i^* = \beta_1 \ln w_i + \beta_2 y_i + \beta_3 x_i + v_i \tag{2.27}$$

where β_1, β_2 and β_3 are unknown parameters of labor supply. Labor supply and wages are now completely described by the triangular system consisting of (2.24) and the following reduced form for desired hours of work:

$$\begin{aligned} h_i^* &= (\beta_1 \gamma_1 + \beta_3) x_i + \beta_1 \gamma_2 q_i + \beta_2 y_i + \beta_1 \eta_i + v_i \\ &= \pi_1 x_i + \pi_2 q_i + \pi_3 y_i + \omega_i \\ &= \pi z_i + \omega_i. \end{aligned} \tag{2.28}$$

2.4.2. Semiparametric estimation

If it can be assumed that v_i and η_i are distributed independently of the explanatory variables x, q and y then semiparametric identification and estimation can take the following simple stepwise procedure.

The π coefficients in (2.28) can be estimated from a standard censored regression estimation procedure. If $g_\omega(\omega)$ describes the density of ω, then the sample likelihood for a random sample of $i = 1, \ldots, n$ individuals is given by

$$L(\pi) = \prod_{i=1}^{n} \{g_\omega(\omega|\pi)\}^{I_i} \left\{1 - \int_{-\pi' z_i} g_\omega(\omega|\pi)\, d\omega\right\}^{1-I_i} \tag{2.29}$$

which is equivalent to the sample likelihood for the Tobit model when ω is homoskedastic normal. Root-n consistent and asymptotically normal estimators of π can be derived under much weaker assumptions on g_ω, see Powell (1987).

Given π, the conditional mean of (2.24) for participants can be used to estimate the wage equation parameters. This is the Heckman (1976, 1979) selectivity framework. Suppose we assume

$$E(\eta_i | I_i > 0) = \lambda_\eta(\pi' z_i), \tag{2.30}$$

then the conditional mean of (2.24) given $I_i > 0$ is simply written as

$$E(\ln w_i | z, I_i > 0) = \gamma_1' x_i + \gamma_2' q_i + \lambda_\eta(\pi' z_i). \tag{2.31}$$

If a joint normal distribution is assumed for v_i and η_i then estimation can follow the two-step selectivity estimation approach developed by Heckman (1979). Alternatively, a $\sqrt{N}$ consistent and asymptotically normal semiparametric estimator can be constructed.

To consider the semiparametric estimator, notice that the conditional expectation of (2.31) for participants given πz_i is

$$E(\ln w_i | \pi' z_i, I_i > 0) = \gamma_1 E(x_i | \pi' z_i) + \gamma_2 E(q_i | \pi' z_i) + \lambda_\eta(\pi' z_i). \tag{2.32}$$

Subtracting this from (2.31) eliminates the $\lambda_\eta(\pi' z_i)$ term yielding

$$\begin{aligned} &E(\ln w_i | z, I_i > 0) - E(\ln w_i | \pi' z_i, I > 0) \\ &\quad = \gamma_1' \big(x_i - E(x_i | \pi' z_i)\big) + \gamma_2' \big(q_i - E(q_i | \pi' z_i)\big). \end{aligned} \tag{2.33}$$

The conditional expectation terms $E(\ln w_i | \pi z_i)$, $E(x_i | \pi z_i)$ and $E(q_i | \pi z_i)$ in (2.33) can then be replaced by their unrestricted Nadaraya–Watson kernel regression estimators.[10]

[10] E.g.

$$\widehat{E(q_i | \pi' z_i)} = \hat{q}^h(\pi z) = \frac{\hat{r}(\pi' z)}{\hat{f}(\pi' z)}$$

The parameters of (2.31) can then be recovered by an instrumental variable regression. Robinson (1988) suggests regressing $\ln w - \widehat{\ln w}^h(\pi z)$ on $x - \hat{x}^h(\pi z)$ and $q - \hat{q}^h(\pi z)$ using $I[\hat{f}(\pi z) > b_N]x$ and $I[\hat{f}(\pi z) > b_N]q$ as the respective instrumental variables, where $I[\hat{f}(\ln x) > b_N]$ is an indicator function that trims out observations for which $\hat{f}(\ln x) < b_N$, for some sequence of trimming constants b_N which tend to zero with the sample size at some appropriate rate. An alternative estimator, due to Powell (1987), is to use $\hat{f}(\pi z).x$ and $\hat{f}(\pi z).q$ as instruments. This effectively removes the random denominators from the kernel regression estimators.

Finally, given the γ_1, γ_2, π_1, π_2 and π_3 parameters, the structural labor supply parameters β_1, β_2 and β_3 can be recovered by minimum distance. In general, these steps can be combined to improve efficiency. Provided a suitable instrumental variable is available, this procedure can also be extended to control for the endogeneity of other income y_i. We consider this in more detail below.

2.5. *Fixed costs*

2.5.1. *A structural model of fixed costs*

Fixed costs imply that participation does not simply follow the corner-solution condition (2.22). Instead participation will depend on the determinants of fixed costs as well as the determinants of h_i^*. For example, suppose there is a fixed monetary cost of working S; this implies that nonlabor income in the budget constraint becomes

$$\begin{aligned} &y - S \quad \text{if } h > 0, \\ &y \qquad\ \ \text{if } h = 0, \end{aligned}$$

and the distribution of S is only partially observable. If we denote utility in work at the optimal hours point by the indirect utility level: $v(w, y, v)$ and utility at $h = 0$ by the direct utility at $h = 0$: $U(Y, 0, v)$, the decision to work follows from

$$v(w, y, v) \geqslant U(Y, 0, v).$$

Note that if $S > \underline{S} \gg 0$ then there will be a discontinuity in the hours distribution at low wages which should reflect itself as a "hole" at the low end of the hours distribution.[11]

in which

$$\hat{r}(\pi z) = \frac{1}{n}\sum_i K_h(\pi z - \pi z_i)q_i \quad \text{and} \quad \hat{f}(\pi z) = \frac{1}{n}\sum_i K_h(\pi z - \pi z_i),$$

where $K_h(\cdot) = h^{-1}k(\cdot/h)$ for some symmetric kernel weight function $k(\cdot)$ which integrates to one. The bandwidth h is assumed to satisfy $h \to 0$ and $nh \to \infty$ as $n \to \infty$. Under standard conditions the estimator is consistent and asymptotically normal, see Härdle and Linton (1994).

[11] This may not be visible since heterogeneity in fixed costs and in unobserved tastes may imply a different position for the discontinuity for different individuals, smoothing out the unconditional distribution. Hence

This model is further developed in Section 4; here we analyze empirical models that are motivated by the presence of fixed costs of work.

Cogan (1981) defines reservation hours h_0 such that

$$U^1(T - h_0, y - S + wh_0, x, v) = U^0(T, y),$$
$$h_0 = h_0(y - S, w, x, v) \geqslant 0 \tag{2.34}$$

and the participation decision becomes

$$\Pr(work) = \Pr\bigl(h > h^0\bigr). \tag{2.35}$$

For any v and η, nonparticipation will occur if fixed costs are sufficiently high: $S > S^*(v, \eta)$.

Suppose we continue to assume wage equation (2.24) and also assume the specification of fixed costs to be

$$S = \theta_1 x + \theta_2 m + s \tag{2.36}$$

where m are a set of variables exclusive to the determination of fixed costs and s represents unobserved heterogeneity in the distribution of fixed costs. In terms of the likelihood contributions we have for the "no work" regime:

$$\ell_0 = \int_{-\infty}^{\infty} \int_{-\infty}^{v^*} \int_{S^*}^{\infty} g(v, \eta, s)\, ds\, dv\, d\eta; \tag{2.37}$$

and for the work regime:

$$\ell_1 = \int_{v^*}^{\infty} \int_{0}^{S^*} g(\varepsilon, v, \eta, s)\, ds\, dv. \tag{2.38}$$

Given some parametric specification of direct (and indirect) utility, all the structural parameters of fixed costs, preferences and wage determination are identified from a likelihood based on the contributions (2.37) and (2.38).

Finally note that if we specify a model on the basis of the indirect utility or cost function we may not have an analytical expression for the direct utility function. Consequently this has to be obtained numerically. One way of doing this is to find the standard reservation wage when hours are zero and the fixed costs have not been incurred. Evaluating the indirect utility function at that reservation wage and nonlabor income then provides us with the utility value of not working. Another important difficulty is then to derive the probability of participation given that the direct utility function at zero hours of work will depend on unobserved heterogeneity both directly and via the reservation wage – hence it is likely to be a highly nonlinear function of the underlying error term. In practice, as we argue later, it may be easier to work with a direct utility specification when we have to deal with such nonconvexities.

looking for such "odd" features in the hours distribution may not be a very good empirical strategy for detecting fixed costs. However such features can be seen in the distribution of relatively homogeneous groups, e.g. single women with no children or single men.

2.5.2. Semiparametric estimation in the fixed costs model

Although the (semiparametric) censored regression approach to the estimation of the hours equation described above is no longer valid in this fixed costs case, a semiparametric procedure applied to hours of work among the participants can be used as an approximation to the fixed costs model. The optimal choice of hours of work among those individuals who decide to join the labor market will have the form

$$\begin{aligned} h_i^* &= \beta_1 \ln w_i + \beta_2(y_i - S_i) + \beta_3 x_i + v_i \\ &= (\beta_1\gamma_1 + \beta_2\theta_1 + \beta_3)x_i + \beta_1\gamma_2 q_i + \beta_2 y_i + \beta_2\theta_2 m_i + \beta_1\eta_i + \beta_2 s_i + v_i \\ &= \tilde{\pi}_1 x_i + \tilde{\pi}_2 q_i + \tilde{\pi}_3 y_i + \tilde{\pi}_4 m_i + u_i \\ &= \tilde{\pi}\tilde{z}_i + u_i \end{aligned} \tag{2.39}$$

where again the β_1, β_2 and β_3 are unknown parameters of labor supply. Labor supply and wages are now completely described by the triangular system consisting of (2.24) and (2.39).

Assume that the participation condition (2.35) can be well approximated by the discrete index model

$$I_i = 1 \quad \Leftrightarrow \quad \phi\tilde{z}_i + e_i > 0 \tag{2.40}$$

where $\tilde{z}_i$ contains all the exogenous variables determining reservation hours, log wages and desired hours of work. The term e_i is a random unobservable whose distribution F_e is normalized up to scale and assumed to be independent of $\tilde{z}_i$. Parameters ϕ will be a convolution of parameters of fixed costs, the wage equation and preferences. They can be identified through the condition

$$E(I_i = 1|\tilde{z}_i) = \int_{-\phi\tilde{z}_i} dF_e(e). \tag{2.41}$$

The ϕ coefficients in (2.40) can be estimated up to scale from a standard binary choice estimation procedure which replaces the censored regression rule (2.21) in this fixed costs model. The sample likelihood for a random sample of $i = 1, \ldots, n$ individuals is given by

$$\Im(\phi) = \prod_{i=1}^{n} \left\{ \int_{-\phi\tilde{z}_i} dF(e) \right\}^{I_i} \left\{ 1 - \int_{-\phi\tilde{z}_i} dF(e) \right\}^{1-I_i} \tag{2.42}$$

which is equivalent to the probit likelihood when e is homoskedastic normal. $\sqrt{N}$ consistent and asymptotically normal estimators of ϕ up to scale can be derived under much weaker index assumptions on f, see Klein and Spady (1993) for example.

Given ϕ, the conditional mean of (2.24) for participants can be used to estimate the wage and hours equation parameters. This is the Heckman (1976, 1979) selectivity framework. Suppose we assume the single index framework

$$E(\eta_i|I_i > 0, \tilde{z}_i) = \lambda_\eta(\phi\tilde{z}_i) \tag{2.43}$$

and

$$E(u_i|I_i > 0, \tilde{z}_i) = \lambda_u(\phi\tilde{z}_i), \tag{2.44}$$

then the conditional mean of (2.24) and (2.39) given $I_i > 0$ are simply written

$$E(\ln w_i|I_i > 0) = \gamma_1 x_i + \gamma_2 q_i + \lambda_\eta(\phi\tilde{z}_i) \tag{2.45}$$

and

$$\begin{aligned} &E(h_i|I_i > 0, \tilde{z}_i) \\ &\quad = (\beta_1\gamma_1 + \beta_2\theta_1 + \beta_3)x_i + \beta_1\gamma_2 q_i + \beta_2 y_i + \beta_2\theta_2 m_i + \lambda_u(\phi\tilde{z}_i), \end{aligned} \tag{2.46}$$

where $\tilde{z}_i$ is taken to include all exogenous variables. If a joint normal distribution is assumed for v_i, η_i and s_i then estimation can follow the two-step selectivity estimation approach developed by Heckman (1979).

Notice that (2.45) and (2.46) together only identify γ_1, γ_2, β_1 and β_2; the parameters of fixed costs and β_3 are not identified without more information on fixed costs. A $\sqrt{N}$ consistent and asymptotically normal semiparametric estimator of these parameters can be constructed from a natural extension of the procedures described above for the censored labor supply model.

For participants we have

$$\begin{aligned} E(h_i|\phi\tilde{z}_i,\ I_i > 0) = {} & \tilde{\pi}_1 E(x_i|\phi\tilde{z}_i,\ I_i > 0) + \tilde{\pi}_2 E(q_i|\phi\tilde{z}_i,\ I_i > 0) \\ & + \tilde{\pi}_3 E(y_i|\phi\tilde{z}_i,\ I_i > 0) + \tilde{\pi}_4 E(m_i|\phi\tilde{z}_i,\ I_i > 0) \\ & + \lambda_u(\phi\tilde{z}_i). \end{aligned} \tag{2.47}$$

The nonparametric term describing the selection of participants can be eliminated as in (2.33) and root-n estimation of the unknown index parameters can also follow the same semiparametric techniques.[12]

Finally, we should note that endogeneity of y_i can be handled in a similar fashion. Suppose a reduced form for y is given by

$$y_i = \vartheta' d_i + \zeta_i; \tag{2.48}$$

since y_i is continuously observed for all individuals, ϑ can be estimated by least squares. Now suppose we also assume that

$$E(u_i|I_i > 0, y_i, \tilde{z}_i) = \delta_y\zeta_i + \lambda_u(\phi\tilde{z}_i). \tag{2.49}$$

Then adding the estimated residual from the regression (2.48) into the selection model (2.46) appropriately corrects for the endogeneity of y_i. This is an important consideration given the consumption-based definition of y_i in the life-cycle consistent specification.

[12] See Newey, Powell and Walker (1990) for some empirical results using semiparametric selection methods.

3. Difference-in-differences, natural experiments and grouping methods

One of the central issues in labor supply is the endogeneity of marginal (post-tax) wages and other incomes. The work incentives facing individuals are usually endogenous. Consider as an example a world with a progressive tax system, as will be examined in detail in the next section. In this case individuals earning more face a higher rate of tax and hence a lower marginal incentive to work. Now take two individuals both of whom have the same pre-tax wage but different tastes for work. The person working longer hours will earn more and will face a higher tax rate, which translates to a lower post-tax marginal wage. In a simple regression framework we would estimate a negative effect of the wage on hours of work since the person with higher hours (because of tastes) will be facing a lower wage. This kind of endogeneity has prompted researchers to seek exogenous sources of variation in policy that resemble experimental situations with a "treatment" group affected by the policy and a "control" or "comparison" group which is unaffected. The impact of incentives is then estimated by comparing the change in hours between the two groups before and after the policy is implemented.

Using this basic idea one can attempt to estimate a "causal" impact of the policy on labor supply, ignoring any structural considerations.[13] Alternatively one can think of the policy changes as an attempt to obtain quasi-experimental conditions for estimating the structural parameters themselves. The former approach attempts to ignore the underlying theory and wishes to go straight to the effects of the particular policy. The latter is after structural parameters that can be used for extrapolation to other policy situations, assuming the theory is a good approximation of reality.

In the following sections we describe this approach to estimating the impact of incentives on labor supply drawing in part from Blundell, Duncan and Meghir (1998). We also discuss the validity of the approach under different circumstances.[14] As one may expect, even the "atheoretical" approach which seeks to estimate the impacts of policy without reference to a model does implicitly make strong assumptions about behavior and/or the environment, and we discuss this. We also discuss conditions under which the quasi-experimental approach, which is a form of instrumental variables, can provide estimates of structural parameters. We go through the difference-in-differences estimator and a more general grouping estimator considering also the effects of selection due to nonparticipation.

3.1. *Difference-in-differences and fixed effects models*

Suppose one is interested in estimating the influence of a policy instrument on an outcome for a group, say outcome h_{it} measuring hours of work or participation. The group consists of individuals $i = 1, \ldots, N$, with these individuals observed over a sample

[13] See Eissa and Liebman (1995) as an example.
[14] See also Blundell and MaCurdy (1999).

horizon $t = 1, 2, \ldots$. Suppose further that a policy instrument of interest changes in a particular period t for only a segment of the individuals. Let δ_{it} be a zero-one indicator that equals unity if the policy change was operative for individual i in period t. Members of the group who experience the policy change react according to a parameter γ. A framework for estimating expressed in terms of a conventional fixed effect model takes the form

$$h_{it} = \gamma\delta_{it} + \eta_i + m_t + \varepsilon_{it} \tag{3.1}$$

where i is a time-invariant effect unique to individual i, m_t is a time effect common to all individuals in period t, and ε_{it} is an individual time-varying error distributed independently across individuals and independently of all η_i and m_t.

The least squares estimator of γ in (3.1), which regresses h_{it} on δ_{it} and a set of individual and time dummy variables, is precisely the difference-in-differences estimator for the impact of the reform. It can be given the interpretation of a causal impact of the reform if $E(\varepsilon_{it}|\eta_i, m_t, \delta_{it}) = 0$. In a heterogeneous response model

$$h_{it} = \gamma_i\delta_{it} + \eta_i + m_t + \varepsilon_{it} \tag{3.2}$$

the least squares dummy variable estimator recovers the average of the response parameters γ_i for those affected by the policy. Since the error term ε_{it} may be correlated both over time and across individuals, this should be taken into account when constructing standard errors.

Now suppose that the policy does not affect everyone in a treatment group, but that the chance of being affected is higher among them ($g = T$) than it is among a control group ($g = C$). The error structure can be more general than above. Consider a specification in which

$$h_{it} = \gamma\delta_{it} + u_{it}, \tag{3.3}$$

where u_{it} represents an individual-level heterogeneity term which may be fixed for that individual over time or may vary over time. Moreover it may be correlated across individuals and may have a cross-section mean that is nonzero. The implicit macro component and the average group characteristics to which the individual belongs may be correlated with δ_{it}. Suppose that limited time-series data is available across individuals, either in terms of repeated cross-sections or as a panel data source. Under the following assumption, and the presumption that the policy is introduced only for one group, the impact of the policy may be identified by using two time periods of data, one before the reform and one after. The assumption we require is that

$$\text{A1:} \quad E[u_{it}|g, t] = a^g + m_t \tag{3.4}$$

which can be interpreted as saying that in the absence of the reform the changes in group means are the same across the two groups. Then with two groups and two time periods the slope coefficient γ can be written as

$$\gamma = \frac{\Delta E(h_{it}|T, t) - \Delta E(h_{it}|C, t)}{\Delta E(\delta_{it}|T, t) - \Delta E(\delta_{it}|C, t)},$$

the difference-in-differences estimator is the sample analog given by

$$\hat{\gamma} = \frac{\Delta \bar{h}_t^T - \Delta \bar{h}_t^C}{\Delta \Pr(\delta_{it} = 1|T,t) - \Delta E(\delta_{it} = 1|C,t)} \tag{3.5}$$

where the "bar" denotes sample average, Δ the first difference and the superscript the group for which first differences are taken. $\hat{\gamma}$ is consistent for γ. This estimator is an instrumental variables estimator with excluded instruments the group–time interactions. If the effect of the treatment is heterogeneous and if the policy does not decrease the chance of obtaining the treatment δ for anyone (monotonicity) then the difference-in-differences estimator above identifies the impact of the policy for those obtaining treatment as a result of the policy [Imbens and Angrist (1994)].

Assumption A1 is very strong indeed. Failure will result if there is a change in group composition of unobservable individual effects over time or if there is a differential impact of macro shocks across groups. Again it will depend critically on the choice of groups which is a key issue in this framework. A1 implies:

(i) time-invariant composition for each group, and
(ii) common time effects across groups.

3.2. Estimating a structural parameter

Here we consider the use of this method in the estimation of a simple labor supply model (ignoring income effects for notational simplicity; we return to this below)

$$h_{it} = \alpha + \beta \ln w_{it} + u_{it}. \tag{3.6}$$

Again u_{it} represents an individual-level heterogeneity term which may be fixed for that individual over time or may vary over time. Moreover it may be correlated across individuals and may have a cross-section mean that is nonzero. This represents the impact of macro shocks to preferences on individual i's labor supply. Both the implicit macro component and the idiosyncratic heterogeneity may be correlated with the log wage ($\ln w_{it}$).

Make the following assumptions:

$$\text{A1:} \quad E[u_{it}|g,t] = a^g + m_t, \tag{3.7}$$

$$\text{A2:} \quad \big[E[\ln w_{it}|g = T,t] - E[\ln w_{it}|g = C,t]\big] - \big[E[\ln w_{it}|g = T,t-1] - E[\ln w_{it}|g = C,t-1]\big] \neq 0. \tag{3.8}$$

Then with two groups and two time periods the slope coefficient β can be written as

$$\beta = \frac{\Delta E(h_{it}|T,t) - \Delta E(h_{it}|C,t)}{\Delta E(\ln w_{it}|T,t) - \Delta E(\ln w_{it}|C,t)}.$$

The difference-in-differences estimator is the sample analog given by

$$\hat{\beta} = \frac{\Delta \bar{h}_t^T - \Delta \bar{h}_t^C}{\Delta \overline{\ln w}_{it}^T - \Delta \overline{\ln w}_{it}^C} \tag{3.9}$$

and is consistent for β.

Assumption A2 is simply analogous to a rank condition and should hold if the groups are chosen to reflect some systematic reason for a differential growth in $\ln w_{it}$ across groups. The choice of groups in this difference-in-differences approach usually reflects some policy change which affects the real wage – a tax change, for example, can be argued to be incident on individuals in one group $i \in [g = T]$ but not on individuals in another $i \in [g = C]$. It is clear, however, that the assumption A1 may be strong in some circumstances. However note the big difference with the previous section. In the previous section the policy was assumed to have no effect on wages of the treatment group relative to the control group; this is the assumption implicit in the fact that we only need to condition on time and group effects. Here we are conditioning on wages and we are adding the assumption from economic theory, that log wages and taxes share the same coefficient. Hence if the policy implicitly affecting incentives changes pre-tax wages as well, this is allowed for; this in itself makes the assumptions underlying the difference-in-differences approach more credible (see more on this below).

This method has some attractive features. It allows for correlated heterogeneity and for general common time effects. Although for many choices of grouping, often precisely those associated with some policy reform, assumption A1 is likely to be invalid, there are possible grouping procedures for estimating labor supply models that are more convincing. This approach is also closely related to the natural experiment or quasi-experimental estimators that typically employ before and after comparisons relating directly to a policy reform.

Before moving on to consider these developments, we first simply outline how this approach can be extended to allow for many groups, for many time periods (or many reforms), for participation and for the inclusion of income terms and other regressors.

3.3. Grouping estimators

Suppose individuals can be categorized in one of a finite number of groups g each sampled for at least two time periods. For any variable x_{it}, define D_x^{gt} as the residual from the following regression

$$E(x_{it}|P_{it}, g, t) = \sum_{g=1}^{G} \zeta_g d_g + \sum_{t=1}^{T} \xi_t d_t + D_x^{gt}, \tag{3.10}$$

where P_{it} indicates that the individual is observed working, that is $P_{it} \equiv \{I_{it} = 1\}$ and where d_g and d_t are group and time dummies respectively. Analogously with A1 and A2 we make assumptions

$$\text{A1.1:} \quad E(u_{it}|P_{it}, g, t) = a_g + m_t, \tag{3.11}$$

$$\text{A2.1:} \quad E\left[D_w^{gt}\right]^2 \neq 0. \tag{3.12}$$

Assumption A1.1 summarizes the exclusion restrictions for identification; it states that the unobserved differences in average labor supply across groups can be summarized by a permanent group effect a_g and an additive time effect m_t. In other words differences

in average labor supply across groups, given the observables, remain unchanged over time. It also says that any self-selection into employment (the conditioning on P_{it}) can be controlled for by group effects and time effects additively. Assumption A2.1 is again equivalent to the rank condition for identification; it states that wages grow differentially across groups; this is because the assumption requires that after we have taken away time and group effects there is still some variance of wages left. For example, if there is a tax reform between two periods, affecting the post-tax wages of the two groups in different ways, and assuming that tax incidence does not fully counteract the effects of the reforms, identification of the wage elasticity will be guaranteed.[15]

With these assumptions we can implement a generalized Wald estimator [see Heckman and Robb (1985)]. Defining the sample counterpart of D_x^{gt} as $\tilde{x}_{gt}$, i.e. the residual from regressing the time-group cell mean on time and group dummies, we can write the estimator as

$$\hat{\beta} = \frac{\sum_g \sum_t [\tilde{h}_{gt}][\widetilde{\ln w}_{gt}] n_{gt}}{\sum_g \sum_t (\widetilde{\ln w}_{gt})^2 n_{gt}} \tag{3.13}$$

where n_{gt} is the number of observations in cell (g, t). The implementation of this estimator is simple; group the data for workers by g and by time and regress by weighted least squares the group average of hours of work on the group average of the log wage, including a set of time dummies and group dummies. An alternative that gives numerically identical results is as follows: regress using OLS the log after-tax wage rate on time dummies interacted with the group dummies, over the sample of workers only and compute the residual from this regression. Then use the original data to regress hours of work on the individual wage, a set of time dummies and group dummies and the wage residual. The t-value on the coefficient of the latter is a test of exogeneity, once the standard errors have been corrected for generated regressor bias and intra-group dependence. It is also important to allow for the possibility of serial correlation and correlation of idiosyncratic shocks across individuals when computing the standard errors.

3.3.1. Controlling for participation

A potential problem with the approach above is that it assumes that the composition effects from changes in participation can be fully accounted for by the additive group and time effects, $a_g + m_t$. First, changes in m_t will cause individuals to enter and leave the labor market. Second, with nonconvexities, a tax policy reform may lead to changes in participation. This will be particularly true if fixed costs are large relative to the nontaxable allowance. The presence of composition effects is equivalent to saying that $E(u_{it}|P_{it}, g, t)$ is some general function of time and group and does not have the additive structure assumed in A1.1.

To control for the possibility that $E(u_{it}|P_{it}, g, t)$ may vary over time requires structural restrictions. A parsimonious specification is to make the assumption of linear

[15] See Bound, Jaeger and Baker (1995) for the implications of weak instruments in empirical models.

conditional expectation. For example, we may extend A1.1 and A2.1 by assuming that

$$\text{A1.2:} \quad E(u_{it}|P_{it}, g, t) = a_g + m_t + \delta\lambda_{gt}, \tag{3.14}$$

$$\text{A2.2:} \quad E\big[D_w^{gt\lambda}\big]^2 \neq 0 \tag{3.15}$$

where λ_{gt} is the inverse Mills' ratio evaluated at $\Phi^{-1}(L_{gt})$, Φ^{-1} being the inverse function of the normal distribution and L_{gt} being the proportion of group g working in period t.[16] Finally $D_w^{gt\lambda}$ is defined by the population regression

$$E(w_{it}|P_{it}, g, t) = \sum_{g=1}^{G} \zeta_g d_g + \sum_{t=1}^{T} \xi_t d_t + \delta_w \lambda_{gt} + D_w^{gt\lambda}. \tag{3.16}$$

Assumption A1.2 models the way that composition changes affect differences in the observed labor supplies across groups. It implies that

$$E(h_{it}|P_{it}, g, t) = \beta E(\ln w_{it}|P_{it}, g, t) + a_g + m_t + \delta\lambda_{gt} \tag{3.17}$$

where all expectations are over workers only. Assumption A2.2 states that wages must vary differentially across groups over time, over and above any observed variation induced by changes in sample composition. We have also implicitly assumed that $E[D_\lambda^{gt}]^2 \neq 0$. If this is not the case, there is no selection bias on the coefficients of interest (here the wage effect) because composition effects can be accounted for by the linear time and group effects. In this case we can use (3.13).

We can now estimate the wage effect using a generalization of (3.13), i.e.

$$\hat{\beta} = \frac{\sum_g \sum_t [\tilde{h}_{gt\lambda}][\widetilde{\ln w}_{gt\lambda}] n_{gt}}{\sum_g \sum_t (\widetilde{\ln w}_{gt\lambda})^2 n_{gt}}. \tag{3.18}$$

As before, this estimator can be implemented using a residual addition technique. We can add an estimate of λ_{gt} as well as the residual of the wage equation estimated on the sample of workers (with no correction for sample selection bias as implied by (3.17)) to an OLS regression of individual hours on individual wages, time dummies and group dummies.

To determine whether (3.18) or (3.13) should best be used we can test the null hypothesis that $E[D_\lambda^{gt}]^2 = 0$ which implies that the group effects a_g and the time effects m_t adequately control for any composition changes (given our choice of groups). If we do not reject this we can use (3.13).

The assumption in A1.2 is worth some discussion. First note that where all regressors are discrete and a full set of interactions are included in the selection equation, use of the normal distribution to compute $\hat{\lambda}_{gt}$ imposes no restrictions. However, the linear conditional expectation assumption implies that a term linear in $\hat{\lambda}_{gt}$ is sufficient to control for selection effects and is potentially restrictive. Using the results in Lee (1984) in general we have that

[16] See Gronau (1974) and Heckman (1974a, 1979).

$$E(u_{it}|P_{it}, g, t) = a_g + m_t + \sum_{k=1}^{K} \delta_k \lambda_{gt}^{(k)} \tag{3.19}$$

where $\lambda_{gt}^{(k)}$ are generalized residuals of order k. The linearity reduces the number of parameters to be estimated and hence the number of periods over which we require exogenous variability in wages. If it is found that $E[D_{\lambda}^{gt}]^2 \neq 0$ then one can experiment by including higher-order generalized residuals after checking that they display sufficient independent variability.

3.3.2. *Income effects*

Income effects are important for labor supply and we need to take them into account for at least two reasons. First, the wage elasticity cannot in general be interpreted as an uncompensated wage elasticity, unless we control for other income. Second, income effects are important if we wish to compute compensated wage elasticities for the purpose of evaluating the welfare effects of tax reforms. It is straightforward to extend the estimator in (3.18) to allow for extra regressors, such as other income. This involves regressing $\tilde{h}_{gt\lambda}$ on $\widetilde{\ln w}_{gt\lambda}$ and $\tilde{y}_{gt\lambda}$ where y is household other income. The rank condition for identification is now more stringent: It requires that the covariance matrix $V = Ez_{gt\lambda}z'_{gt\lambda}$ is full rank, where $z_{gt\lambda} = [D_w^{gt\lambda}, D_y^{gt\lambda}]'$.

This is equivalent to requiring that the matrix of coefficients on the excluded exogenous variables in the reduced forms of log wage and other income, after taking account of composition effects, is rank 2. A necessary but not sufficient condition for this to be true is that these coefficients are nonzero in each of the reduced forms – i.e. that $E(D_w^{gt\lambda})^2$ and $E(D_y^{gt\lambda})^2$ are nonzero. As before if we accept the hypothesis that $E(D_{\lambda}^{gt})^2 = 0$ we need to consider whether the rank of $V^* = Ez_{gt}^* z_{gt}^{*\prime}$ is two, where $z_{gt}^* = [D_w^{gt}, D_y^{gt}]'$. In this case we estimate the model using the sample counterparts of z_{gt}^* as regressors.[17]

3.4. *The difference-in-differences estimator and behavioral responses*

As we have seen the simplest implementation of the difference-in-differences approach simply includes a policy reform dummy. This avoids directly specifying a structural model in the sense that the effect of the policy is not tied to a wage or income effect. The idea is that the policy should be evaluated directly without the intermediation of an economic model.

Suppose again there are simply two periods and two groups. Suppose the policy reform is a tax change in which τ is the change in the marginal tax rate for the treatment group. The *natural experiment* approach simply includes a policy dummy

[17] Blundell, Duncand and Meghir use the changes in the distribution of wages, as documented in Gosling, Machin and Meghir (2000) as a source of exogenous variation in wages.

$\delta_t^g = 1\{g = T,\ t = A\}$ in the hours regression

$$h_i = \alpha + \beta\delta_t^g + \zeta_{it}. \tag{3.20}$$

The quasi-experimental estimator in this case is just the difference-in-differences estimator applied to (3.20).

To interpret this estimator suppose the hours equation has the simple form (3.6). Suppose that pre- and post-reform wages are defined by

	Before reform	*After reform*
$i \in$ Treated	$\ln w_{iB}$	$\ln\big((1-\tau)w_{iA}\big)$
$i \in$ Control	$\ln w_{iB}$	$\ln w_{iA}$

Assuming A1 and A2, taking group means, we find

$$\overline{h}_t^g = \alpha + \beta \ln(1-\tau)\delta_t^g + \beta\overline{\ln w}_t^g + a^g + m_t. \tag{3.21}$$

If $\delta_t^g = 1\{g = T,\ t = A\}$ is all that is included in the regression then the difference-in-differences estimator will only recover β if log wages have the group and common time effect form

$$\overline{\ln w}_t^g = \tilde{a}^g + \tilde{m}_t. \tag{3.22}$$

This seems a particularly strong assumption given empirical knowledge about the differential trends in wage rates across different groups in the economy. Clearly, the cost of including simply the policy reform dummy $\delta_t^g = 1\{g = T,\ t = A\}$ alone is that the common time effects and time-invariant composition effects assumptions become even more difficult to satisfy.

4. Estimation with nonlinear budget constraints

A problem encountered in many analyses of consumption and labor supply involves the presence of intricate nonlinearities in budget sets arising from wages and prices that vary as a function of quantities. Tax and welfare programs constitute a prominent source of such functional relationships in analyses of labor supply, for these programs induce net wages to vary with the number of hours worked even when the gross wages remain constant. Hedonic environments and price schedules dependent upon quantities give rise to comparable sources of distortions in budget sets in many consumption settings.

To address the issues encountered with nonlinear budget sets, there has been steady expansion in the use of sophisticated statistical models characterizing distributions of discrete-continuous variables that jointly describe both interior choices and corner solutions in demand systems. These models offer a natural framework for capturing irregularities in budget constraints, including those induced by the institutional features of tax and welfare programs.[18]

[18] Some of the references include Heckman (1974c), Arrufat and Zabalza (1986), Keane and Moffitt (1998), Hausman (1980, 1981, 1985a, 1985b), Moffitt (1983, 1986), Blomquist (1983, 1996), Blomquist and Hansson-Brusewitz (1990), Blomquist and Newey (2002), MaCurdy, Green and Paarsch (1990), MaCurdy (1992).

This section briefly describes approaches for estimating models incorporating such features, keeping the context general enough to illustrate how these models can readily accommodate a wide variety of nonlinearities in price and wage structures. The discussion begins with a brief overview of the methods implemented to model budget constraints involving nonlinearities, and then goes on to survey instrumental variable procedures applied in the literature to estimate behavioral relationships in the presence of such constraints. We summarize the general approach for using maximum likelihood procedures to estimate the more sophisticated variants of these models with either convex or nonconvex budget sets. We provide simple illustrations of maximum likelihood methods to estimate familiar specifications of labor supply with convex constraints. We outline why the implementation of maximum likelihood procedures imposes interesting and important restrictions on behavioral parameters in the presence of nonlinear budget constraints. We then integrate the analysis of nonparticipation into our analysis of nonlinear budget constraints and discuss estimation when the availability of welfare programs affects the shapes of budget sets, which induces not only nonconvexities but also opportunities for participating in multiple programs. Finally, we consider computational simplifications adopted in the literature to render maximum likelihood estimation feasible.

4.1. Modeling nonlinear features of budget constraints

A general formulation for the economic problem considered in the subsequent discussion specifies an agent as solving the following optimization problem:

$$\text{Max} \quad U(c, h, z, \nu) \quad \text{subject to} \quad b(c, h, W, Y) = 0 \tag{4.1}$$

where $U(\,)$ delineates the utility function, c and h measure consumption and hours of work, the quantities z and ν represent respectively the observed and unobserved factors influencing choices beyond those incorporated in budget sets, and the function $b(\,)$ specifies the budget constraint with W and Y designating the real gross wage per hour and nonlabor income (note that we use upper case to distinguish from marginal wage and virtual nonlabor income). For the moment, we restrict the economic framework to be static and the quantities c and h to be single goods rather than multidimensional vectors. In many applications the budget function, b, is not differentiable, and in some it is not even continuous.

For the familiar linear specification of the budget constraint, b takes the form

$$b(c, h, W, Y) = Wh + Y - c. \tag{4.2}$$

Solving (4.1) for this form of b yields the following labor supply and consumption functions:

$$\begin{aligned} h &= \ell(W, Y, z, \nu), \\ c &= c(W, Y, z, \nu), \end{aligned} \tag{4.3}$$

which correspond to the standard demand functions for nonmarket time (i.e., leisure) and consumption. (The subsequent analysis often suppresses the z argument in the functions $U(\,)$, $\ell(\,)$ and $c(\,)$ to simplify notation.)

Another popular specification of $b(\,)$ incorporates income or sales taxes in characterizing choices, with the budget constraint written as some variant of

$$b(c, h, W, Y) = Wh + Y - c - \tau(Wh, Y), \tag{4.4}$$

where the function $\tau(\,)$ gives the amount paid in taxes. This formulation for b admits different tax rates on earnings (Wh) and nonlabor income (Y). If these income sources are instead taxed the same, then (4.4) further simplifies to

$$b(c, h, W, Y) = Wh + Y - c - \tau(I), \tag{4.5}$$

where tax payments $\tau(I) = \tau(I(h))$ where $I(h) =$ taxable income $= Wh + Y - D$ with D designating allowable deductions. Different marginal tax rates in the various income brackets combined with the existence of nonlabor income create inherent nonlinearities in budget sets.

The literature relies on two approaches for modeling nonlinearities in budget sets: piecewise-linear characterizations and smooth differentiable functions. To illustrate these approaches, the subsequent discussion principally focuses on the income-tax formulation of b given by (4.4) and (4.5) to illustrate central concepts.

4.1.1. Piecewise linear constraints

As a simple characterization of piecewise budget sets, Figure 4.1 shows a hypothetical budget constraint for an individual faced with a typical progressive income tax schedule defined by a series of income brackets. In this diagram, h denotes hours of work, and the vertical axis measures total after-tax income or the consumption of market goods. The budget constraint is composed of several segments corresponding to the different marginal tax rates that an individual faces. In particular, he faces a tax rate of t_1 between H_0 hours and H_1 hours (segment 1 of his constraint) and tax rates of t_2 and t_3 respectively in the intervals (H_1, H_2) and $(H_2, \overline{H})$ (segments 2 and 3 in the figure). Thus, with the variable W denoting the individual's gross wage rate, the net wages associated with each segment are: $w_1 = (1 - t_1)W$ for segment 1; $w_2 = (1 - t_2)W$ for segment 2; and $w_3 = (1 - t_3)W$ for segment 3. Also, each segment has associated with it a virtual income (i.e., income associated with a linear extrapolation of the budget constraint) calculated as:

$$\begin{aligned} &y_1 = Y - \tau(0, Y); \\ &y_j = y_{j-2} + W(t_{j-2} - t_j)\overline{h}_{j-1} \quad \text{for } j = 2, 3, \ldots. \end{aligned} \tag{4.6}$$

So, $y_3 = y_1 + (w_1 - w_3)H_2$ and similarly for y. Changes in tax brackets create the kink points at H_1 and H_2.

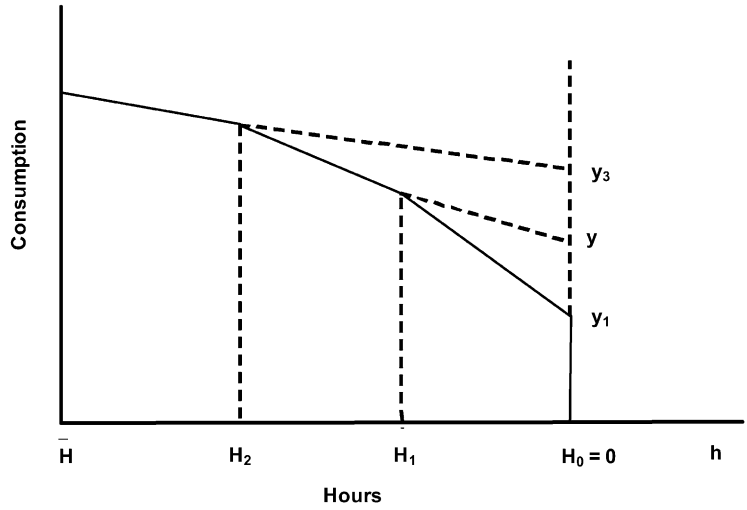

Figure 4.1. Budget constraint with income taxes.

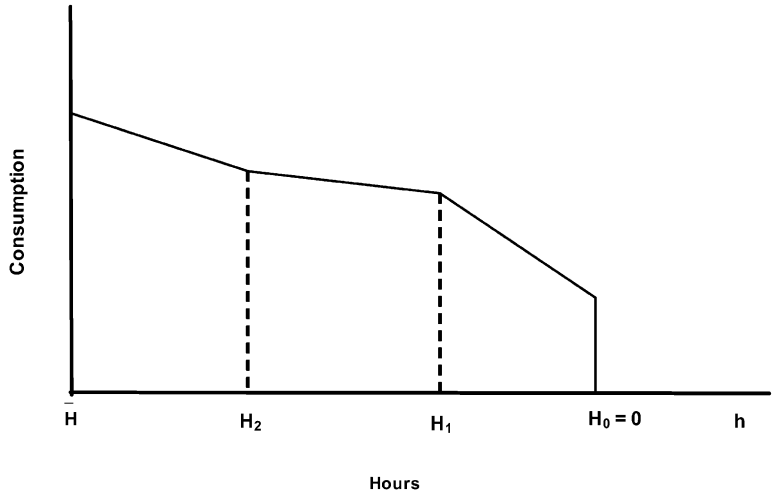

Figure 4.2. Budget constraint with EITC.

Figure 4.2 illustrates stylized features of a budget constraint modified to incorporate an earned income tax credit (EITC) in conjunction with an income tax,[19] and Figure 4.3

[19] An earned income tax credit (EITC) constitutes a negative income tax scheme, which induces two kinks in a person's constraint in the simplest case: one where the proportional credit reached its maximum (H_1 in

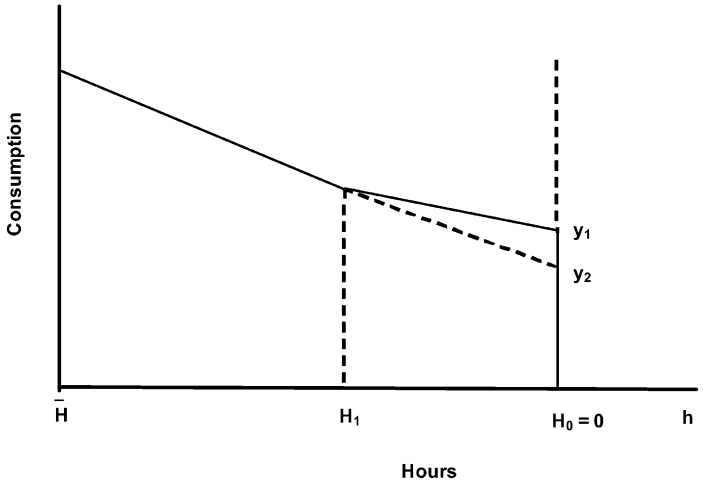

Figure 4.3. Budget constraint with welfare.

shows a prototype budget set induced by a conventional welfare program (or social security tax).[20] In Figure 4.2, the EITC increases benefits until an individual reaches h_1 hours of work, and then benefits decline until hours attain h_2 when the regular income tax schedule applies. In Figure 4.3, welfare benefits start at $y_1 - y_2$ when a family does not work, they steadily decline as the family increases its hours of work until its earnings reach the value implied at H_1 hours when the family becomes disqualified for welfare. Each of these low-income support programs introduces regressive features in the overall tax schedule faced by a family, which in turn induces nonconvex portions in the budget sets.

In real world applications of piecewise budget constraints, the combination of various tax and public assistance programs faced by families implies budget sets have two noteworthy features. First, the constraint faced by a typical individual includes a large number of different rates. Translated into the hours-consumption space, this implies a large number of kink points in the budget constraint. Second, for most individuals the tax schedule contains nonconvex portions, arising from four potential sources. The first arises from the EITC program, as illustrated in Figure 4.2. A second source arises if a

Figure 4.2), and one at the break-even point where the credit was fully taxed away (H_2 in the figure). The tax rates associated with the first two segments are t_A, which is negative, and t_B, which is positive. Thereafter, the EITC imposed no further tax.

[20] A welfare program pays a family some level of benefits at zero hours of work, and then "taxes" this nonlabor income at some benefit reduction rate until all benefits are gone. Figure 4.3 assumes a proportional benefit reduction rate applies on earnings until benefits decline to zero, after which the family pays normal income tax which here too is assumed to be at a proportional rate. Thus, Figure 4.3 shows a constraint with a single interior kink (given by H_1 in the figure) corresponding to the level of earning when welfare benefits first become zero. The Social Security system induces a similar effect on the budget constraint.

worker's family participates in any welfare program, wherein nonconvexities arise as benefits are withdrawn as earnings increase as illustrated in Figure 4.3. Third, social security taxes phase out after a fixed level of earnings, so they too induce a budget set similar in structure to that given by Figure 4.3. Finally, the standard deduction present in most income tax programs, wherein no taxes are paid on sufficiently low levels of income, creates yet another source of regressivity in the tax schedule and corresponding nonconvexities in the budget constraint.

4.1.2. Constructing differentiable constraints

Several approaches exist for approximating the piecewise-linear tax schedules by a differentiable function. A convenient method for constructing this function is to fit the marginal tax rate schedule – a step function – by a differentiable function. This approximation must itself be easily integrable to obtain a simple closed form for the tax function.

An elementary candidate for constructing a differentiable approximation that can be made as close as one desires to the piecewise-linear tax schedule has been applied in MaCurdy, Green and Paarsch (1990). To understand the nature of the approximation, return to Figure 4.1. One can represent the underlying schedule as follows:

$$\begin{aligned} \tau_e(Wh, Y) &= t_1 \quad \text{from } I(H_0) \text{ to } I(H_1) \\ &= t_2 \quad \text{from } I(H_1) \text{ to } I(H_2) \\ &= t_3 \quad \text{above } I(H_2), \end{aligned} \tag{4.7}$$

where $\tau_e(Wh, Y) =$ the marginal tax rate on earnings,

$I(h) =$ taxable income at h hours of work, and

$t_i =$ marginal tax rate, $i = 1, 2, 3$.

For expositional simplicity, suppose that $t_1 = 0$. Consider the following approximation of this schedule:

$$\hat{\tau}_e(Wh, Y) = t_2\big\{\Phi_1\big(I(h)\big) - \Phi_2\big(I(h)\big)\big\} + t_3\Phi_2\big(I(h)\big). \tag{4.8}$$

This formulation for the marginal tax rate switches among three flat lines at the heights t_1 $(= 0)$, t_2 and t_3. The weight functions $\Phi_i(I(h))$ determine the rate at which the shift occurs from one line to another, along with the points at which the switches take place. Candidate weight functions are given by $\Phi_i(I(h)) =$ the cumulative distribution function with mean y_i and variance σ_i^2, $i = 1, 3$. The middle segment of the tax schedule has height t_3 and runs from taxable income $I(H_1)$ to $I(H_2)$. To capture this feature, parameterize $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$ with means $y_1 = I(H_1)$ and $y_3 = I(H_2)$, respectively, with both variances set small. The first distribution function, $\Phi_1(\cdot)$ takes a value close to zero for taxable income levels below $I(H_1)$ and then switches quickly to take a value of one for higher values. Similarly, $\Phi_2(\cdot)$ takes a value of zero until near $I(H_2)$ and one

thereafter. The difference between the two equals zero until $I(H_1)$, one from $I(H_1)$ to $I(H_2)$ and zero thereafter. Thus, the difference takes a value of one just over the range where t_2 is relevant. Notice that we can control when that value of one begins and ends by adjusting the values y and y_3. Also, one can control how quickly this branch of the estimated schedule turns on and off by adjusting the variances of the cumulative distribution functions, trading off a more gradual, smoother transition against more precision. In general, adjusting the mean and variance parameters allows one to fit each segment of a schedule virtually exactly, switch quickly between segments, and still maintain differentiability at the switch points.

A generalization of this approximation takes the form

$$\hat{\tau}_e(Wh, Y) = \sum_{i=1,3,\ldots} \big[\Phi_{i-2}\big(I(h)\big) - \Phi_i\big(I(h)\big)\big] t_i\big(I(h)\big) \tag{4.9}$$

where the functions $t_i(I(h))$ permit tax schedules to be nonconstant functions of taxable income within brackets. With the Φ_i denoting many cdfs associated with conventional continuously distributed distributions, function (4.9) yields closed form solutions when it is either integrated or differentiated.[21] Integrating (4.9) yields a formulation for the budget constraint $b(c, h, W, Y)$. The resulting approximation can be made to look arbitrarily close to the budget set boundary drawn in Figure 4.1, 4.2 or 4.3, except that the kink points are rounded.

Formula (4.9) can be extended to approximate virtually any specification of $b(c, h, W, Y)$. One can readily allow for distinct relationships describing the derivatives for each of the arguments of this function, and nonconvexities in budget sets cause no particular problems.

4.2. *Simple characterizations of labor supply and consumption with differentiable constraints*

A useful solution exists for the hours-of-work and consumption choices associated with utility maximization when budget constraints form a set with a twice-differentiable

[21] Total taxes are given by $\tau(I) = \int \tau'(I)\, dI$. The following relations enable one to calculate an explicit form for $\tau(X)$:

$$\int \Phi\, dI = I\Phi + \varphi,$$

$$\int I\Phi\, dI = \frac{1}{2}I^2\Phi - \frac{1}{2}\Phi + \frac{1}{2}I\varphi,$$

$$\int I^2\Phi\, dI = \frac{1}{3}I^3\Phi + \frac{2}{3}\varphi + \frac{1}{3}I^2\varphi,$$

$$\int I^3\Phi\, dI = \frac{1}{4}I^4\Phi - \frac{3}{4}\Phi + \frac{3}{4}I\Phi + \frac{1}{4}I^3\varphi.$$

In this expression, Φ refers to any Φ_i's, and φ designates the density function associated with Φ_i.

boundary. Specify the marginal wage rate as

$$\omega = \omega(h) = b_h(c, h, W, Y) = b_h \tag{4.10}$$

and "virtual" income as

$$y = y(h) \quad \text{which solves the equation } b(hb_h + y, h, W, Y) = 0. \tag{4.11}$$

This solution for y satisfies

$$y = y(h) = c - \omega h.$$

For the familiar specification given by $b(c, h, W, Y) = Wh + Y - c - \tau(Wh + Y)$ with the function τ constituting the amount paid in taxes at before-tax income $Wh + Y$, the expressions for marginal wage and virtual income y simplify to

$$\begin{aligned} \omega &= \omega(h) = (1 - \tau')W, \\ y &= y(h) = Wh + Y - \omega h - \tau = Y + \tau' Wh - \tau, \end{aligned} \tag{4.12}$$

where τ and τ' (the derivative of the tax function with respect to income) are evaluated at income level $I = I(h) = Y + Wh$ which directly depends on the value of h.

Utility maximization implies solutions for hours of work and consumption that obey the implicit equations:

$$\begin{aligned} h &= \ell(\omega, y, z, v) = \ell\big(\omega(h), y(h), z, v\big), \\ c &= c(\omega, y, z, v) = c\big(\omega(h), y(h), z, v\big), \end{aligned} \tag{4.13}$$

where ℓ and c represent the same conventional forms for labor supply and consumption demand functions given by (4.3). Figures 4.1 and 4.3 illustrate this representation of the solution for optimal hours of work and consumption. The characterization portrays an individual as facing a linear budget constraint in the presence of nonlinear tax programs. This linear constraint is constructed in a way to make it tangent to the actual nonlinear opportunity set at the optimal solution for hours of work. The implied slope of this linearized constraint is $\omega(h)$ and the corresponding value of virtual income is $y(h)$.

Relationships (4.13) constitute structural equations that determine hours of work and consumption. By applying the Implicit Function Theorem to specification (4.13), we can solve this implicit equation for h in terms of W, Y, and other variables and parameters entering the functions b and U. This operation produces the labor supply and consumption functions applicable with general forms of nonlinear budget sets.

4.3. Instrumental variable estimation

The inclusion of taxes provides an additional reason for allowing for the endogeneity of (after-tax) wages and other income. Writing the labor supply function as

$$h = \ell\big(\omega(h), y(h), z, v\big) = \ell^*\big(\omega(h), y(h), z\big) + v \tag{4.14}$$

makes the point explicitly. The instrumental variable approach described earlier can be applied as well as the grouping method (which of course is just an application of IV). The implementation of IV procedures imposes no parametric restrictions and it allows one to consider a wide variety of exogeneity assumptions. The fact that the error term does not interact with the wage and other income is critical for the interpretation of IV as identifying the structural parameters of the model.

4.3.1. Including measurement error

In many data sets there are serious suspicions that hours of work and wages are reported with error. This issue acquires added importance when we are dealing with nonlinear tax schedules since this creates a problem of observability of the correct tax rate, which is the reason we introduce the issue here.

Suppose H denotes measured hours of work and that the function

$$H = H(h, \varepsilon) \tag{4.15}$$

relates H to actual hours h and to a randomly distributed error ε. Typically, analyses presume that the state $h = 0$ is observed without error.

Measurement errors in hours often induce errors in observed wage rates since researchers construct wages by dividing total labor earnings, E, by hours worked in the period. Whereas $W = E/h$ defines the true hourly wage rate, $\widetilde{W} = E/H$ designates the data available on wages. Measured wages $\widetilde{W}$ are contaminated by reporting errors even when E provides accurate quantities for each individual's total labor earnings and wages are indeed constant for hours worked over the period. This formulation presumes a reciprocal relation in the measurement error linking data on hours and wages. More generally, suppose $\widetilde{W}$ links to the true wage rate according to the relationship

$$\widetilde{W} = \widetilde{W}(W, h, \varepsilon). \tag{4.16}$$

In the reciprocal measurement error example, $\widetilde{W} = W/H(h, \varepsilon)$ where $H(h, \varepsilon)$ comes from (4.15).

The presence of measurement errors in hours typically invalidates use of nonlinear IV procedures to estimate the structural labor supply equation given by (4.14). Expressing this equation in terms of H rather than h involves merely substituting (4.15) into (4.14); and if measurement error also carries over to wages, then substitutions must be made for wages as well. These replacements typically result in a variant of structural equation (4.14) that cannot be transformed into a form that is linear in disturbances. Measurement errors in hours invariably render the marginal tax rate unobservable, which in turn makes both the marginal wage ($\omega(h)$) and virtual income ($y(h)$) also unobservable. Sophisticated adjustments must be included to account for such factors. These complications motivate many researchers to turn to maximum likelihood procedures to estimate hours-of-work specifications as we do below. However with some additional assumptions IV procedures are still possible, at least when the issue of censoring does not arise.

Suppose measurement error is of the multiplicative kind

$$H = H(h, \varepsilon) = he^{\varepsilon} \quad \text{with } \widetilde{W} = E/H. \tag{4.17}$$

In the presence of such error, specifications can also be found that allow for use of IV procedures to estimate substitution and income parameters. Incorporating the multiplicative measurement error model (4.17) into the semilog functional form of labor supply given in relation (2.8) yields the empirical specification:

$$H = \bar{u} + Z\gamma + \alpha \ln \omega^m + \beta y + u \tag{4.18}$$

where

$$\begin{aligned}
&\ln \omega^m = \ln(E/H) + \ln(1 - \tau'),\\
&\tilde{y} = y - \alpha\sigma_{\varepsilon}^2/2,\\
&u = v + \alpha\big(\varepsilon - E(\varepsilon)\big) + (H - h) = v + \alpha\big(\varepsilon - E(\varepsilon)\big) + h\big(e^{\varepsilon} - 1\big).
\end{aligned}$$

The disturbance u possesses a zero mean since $E(e^{\varepsilon}) = 1$. Virtual income $y(h)$ and the marginal tax rate τ' are not contaminated by measurement error because they are only functions of Y and $hW = H\widetilde{W}$, quantities which are both perfectly observed (by assumption). The variable $\ln \omega^m$ represents the natural logarithm of the after-tax wage rate evaluated at observed hours, which differs from the actual marginal wage due to the presence of reporting error in hours. Assuming the error ε is distributed independently of all endogenous components determining h, including the heterogeneity disturbance v, the instrumental variables X applicable for estimation of the original specification can also serve as the instrumental variables in estimating the coefficients of (4.18) by familiar IV methods.

4.3.2. Sources of unobservables in budget sets

An important class of models not widely recognized in the literature involves budget constraints that vary across individuals in ways that depend on unobserved factors. The modification required in the above analysis to account for such factors replaces budget function $b(\,)$ appearing in (4.1) by

$$b(c, h, W, Y, z, \xi) = 0. \tag{4.19}$$

The quantity z captures the influence of measured characteristics on budget sets. Classic examples include family characteristics that alter the form of the tax function relevant for families. The error component ξ represents unobserved factors shifting budget sets. Classic examples here include unmeasured components of fixed costs, prices, and elements determining tax obligations.

The presence of ξ in $b(\,)$ typically renders IV methods inappropriate for estimating parameters of the labor supply function ℓ. The usual problem comes about since structural variants of ℓ cannot be found that are linear in disturbances, and this is especially true when nonlinearities exist in tax schedules. When ξ appears as a component of $b(\,)$,

researchers typically rely on the maximum likelihood methods summarized in the subsequent discussion to conduct estimation of behavioral models of hours of work and consumption.

4.3.3. Complications of IV estimation with piecewise-linear constraints

Naive application of instrumental variable methods with piecewise-linear budget constraints generally produces inconsistent estimates of behavioral parameters, even ignoring the potential presence of measurement error. Section 4.6 below presents the structural specification – see (4.69) – implied for hours of work when Figure 4.1 designates the budget set and everyone works. As noted in Section 4.1, this budget set is convex and consists of three segments. Inspection of structural equation (4.69) reveals that the structural error is $\sum_{j=1,2,3} d_j v$ where d_j represents an indicator variable signifying whether an individual selects segment $j = 1, 2$, or 3. If the individual occupies any kink, then $\sum_{j=1,2,3} d_j v = 0$. Suppose X includes the set of instrumental variables presumed to satisfy $E(v|X) = 0$. The corresponding conditional expectation of the structural error implied by Equation (4.69) is $\sum_{j=1,2,3} \Pr(d_j|X) E(v|d_j = 1, X)$. This expectation is typically not zero, a condition required to implement IV techniques. To use IV procedures in the estimation of Equation (4.69) necessitates the inclusion of sample selection terms adjusting for the nonzero expectation of $\sum_{j=1,2,3} d_j v$.

4.3.4. Nonparticipation and missing wages

In the earlier sections we discussed how the estimation approach needs to be generalized so as to allow for nonparticipation and for missing wages, which present further complications for estimation. We argued that standard instrumental variables are not appropriate in this context. We now turn to maximum likelihood estimation which we set up to deal with the problems introduced above, namely nonlinear taxes, measurement error, missing and/or endogenous wages and other income and of course nonparticipation.

4.4. Maximum likelihood estimation: A general representation

The instrumental variable estimator, developed in the last section, required exclusion restrictions to consistently estimate the parameters of the labor supply and consumption models involving nonlinear budget sets. In contrast, maximum likelihood estimation exploits the precise structure of the budget constraint and need not rely on exclusion restrictions to identify parameters. Even though marginal wages and virtual incomes are endogenous, nonlinearities introduced through distributional assumptions provide a valuable source of identification. However, exclusion restrictions are only avoided in this approach if gross wages and incomes are assumed to be exogenous and in many applications of maximum likelihood researchers also impose stringent distributional and independence assumptions on sources of errors capturing heterogeneity and measurement error. Nonetheless, one can entertain a wide array of nonlinearities in budget sets

and decision processes, along with rich specifications for heterogeneity and mismeasurement of variables.

The following discussion begins with a general presentation describing the application of maximum likelihood methods in hours of work and consumption analyses allowing for flexible distributional assumptions and intricate forms of nonlinearities in both preferences and budget constraints. This analysis draws heavily upon Appendix A. Later subsections cover simple illustration of techniques, many of which have been implemented in the empirical literature.

4.4.1. Dividing the budget constraint into sets

Irrespective of whether one considers differentiable or piecewise-linear formulations for budget constraints, the essential idea underlying development of likelihood functions in the presence of nonlinear constraints involves defining a set of "states of the world". Each state designates a particular segment of the budget set boundary, with states being mutually exclusive and states jointly covering all parts of budget constraints. One interprets individuals as being endowed with a set of attributes determining their tastes, resources and constraints, with these attributes viewed as random variables continuously distributed across the population. Based on the realizations of these variables, an individual selects consumption and hours of work to optimize utility.

Regarding the distribution of these variables in the previous discussion, suppose unobserved heterogeneity influencing preferences, ν, the unmeasured factors determining wages, η, and the unobservables incorporating budget sets, ξ, possess the following joint density:

$$\varphi(\upsilon, \eta, \xi) \equiv \varphi(\upsilon, \eta, \xi | X) \quad \text{for } (\upsilon, \eta, \xi) \in \Omega. \tag{4.20}$$

When errors, ε, contaminate the measurement of hours, the relevant joint distribution becomes:

$$\varphi(\upsilon, \eta, \xi, \varepsilon) \equiv \varphi(\upsilon, \eta, \xi, \varepsilon | X) \quad \text{for } (\upsilon, \eta, \xi, \varepsilon) \in \Omega. \tag{4.21}$$

Both these expressions admit conditioning on a set of exogenous variables X, but the subsequent analysis suppresses X to simplify the notation. The set Ω designates the domain of these random variables.

In this setting, n states of the world can occur. The discrete random variable δ_i signifies whether state i happens, with $\delta_i = 1$ indicating realization of state i and $\delta_i = 0$ implying that some state other than i occurred. A state refers to locations on boundaries of budget sets, to be explained further below. Consequently, the value of δ_i depends on where (υ, η, ξ) falls in its domain determined by the rule:

$$\delta_i = \begin{cases} 1 & \text{if } (\upsilon, \eta, \xi) \in \Omega_i, \\ 0 & \text{otherwise,} \end{cases} \tag{4.22}$$

where the set Ω_i constitutes that subset of the sample space Ω for which utility maximization yields a solution for consumption and hours that lies within the $\delta_i = 1$

portion of the budget. The mutually exclusive and exhaustive feature of the sets Ω_i for $i = 1, \ldots, n$ implies $\bigcup_{i=1}^{n} \Omega_i = \Omega$ and $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$.

A central requirement invoked in dividing a budget constraint into its various sections involves ensuring that unique solutions exist for c and h for any $(v, \eta, \xi) \in \Omega_i$. Consumption and hours of work may take on discrete values when $(v, \eta, \xi) \in \Omega_i$. Alternatively, there may be a continuous mapping relating c and h to (v, η, ξ) within the set Ω_i, but inverses must exist for the consumption and labor supply functions

$$h = \ell(\omega, y, z, v) = \ell\big(\omega(h), y(h), z, v\big),$$
$$c = c(\omega, y, z, v) = c\big(\omega(h), y(h), z, v\big)$$

expressed in terms of components of v. (These functions correspond directly to those in (4.13) except that marginal wage ω and virtual income y now are functions of ξ, the unobservable components of b.) Considering the labor supply function ℓ, this requirement implies existence of the inverse function

$$v = v^h\big(h, \omega(h), y(h), z\big) \equiv \ell^{-1}\big(h, \omega(h), y(h), z\big) \tag{4.23}$$

for values of v within the set Ω_i. If v is in fact multidimensional (i.e., v' is a vector), then an inverse must exist of the form

$$v_1 = v^h\big(h, \omega(h), y(h), z, v_2'\big) \equiv \ell^{-1}\big(h, \omega(h), y(h), z, v_2\big) \tag{4.24}$$

for some decomposition $v' = (v_1, v_2')$.

Division of the budget constraint into the events $\delta_i = 1$ for $i = 1, \ldots, n$ generally creates two varieties of sets: First, differentiable segments of the budget constraint over which consumption and hours vary continuously in response to variation in preferences and constraint variables; second, kink points at which consumption and hours of work take fixed discrete values implied by the location of the kink.

4.4.2. Maximum utility determines placement on budget constraint

The portion of a budget constraint selected by an individual depends on the level of utility assigned to this state. The following discussion first characterizes maximum utility attainable on differentiable segments, and then considers evaluations at kink points.

For the differentiable segments of the constraint, utility is determined by function

$$\begin{aligned} V &= U\big(c(\omega, y, z, \nu), \ell(\omega, y, z, \nu), z, \nu\big) \\ &= U\big(c\big(\omega(h, \xi), y(h, \xi), z, \nu, \xi\big), \ell\big(\omega(h, \xi), y(h, \xi), z, \nu, \xi\big), z, \nu\big) \\ &\equiv V\big(\omega(h, \xi), y(h, \xi), z, v\big) \\ &= V(\omega, y, z, v) \end{aligned} \tag{4.25}$$

evaluated at optimal points in the specified set. The function $V(W, Y, z, v)$ represents the conventional indirect utility function associated with maximizing $U(C, h, z, \nu)$ in (4.1) subject to the linear form of the budget constraint given by (4.2). Roy's identity

specifies that the labor supply function ℓ can be written as

$$\ell(\omega, y, z, \upsilon) \equiv \frac{V_\omega(\omega, y, z, \upsilon)}{V_y(\omega, y, z, \upsilon)} \tag{4.26}$$

with V_ω and V_y denoting the partial derivatives of V. Suppose the interval $(\overline{h}_{i-1}, \overline{h}_{i+1})$ identifies the differential segment under consideration. The subsequent discussion refers to this segment as state i. Then the utility assigned to state i corresponds to the maximum value of V achievable for hours falling in the interval $(\overline{h}_{i-1}, \overline{h}_{i+1})$.

Difficulty in determining the achievable value of V depends on characteristics of the budget function $b(c, h, W, Y)$. For the most general specifications of b, inspection of relations (4.10) and (4.11) defining ω and y reveals that each depends on both c and h through the derivative b_h. If utility maximization occurs at an interior point of $(\overline{h}_{i-1}, \overline{h}_{i+1})$ given the realization of (υ, η, ξ), then the implied values of c and h solve the system

$$\begin{aligned} &h = \ell(\omega, y, z, \upsilon) \in (\overline{h}_{i-1}, \overline{h}_{i+1}), \\ &b(c, h, W, Y, \xi) = 0. \end{aligned} \tag{4.27}$$

Consequently, the maximum utility attainable on the interval $(\overline{h}_{i-1}, \overline{h}_{i+1})$ is V (or U) evaluated at these solutions for c and h. Define this maximum utility as $V_{(i)}$, where the (i) subscript on V signifies utility assigned to state i. If one extends state i to include either of the exterior points $\overline{h}_{i-1}$ or $\overline{h}_{i+1}$, and uniqueness and differentiability continue to hold at these points, then the above procedure still applies in assigning a value for $V_{(i)}$. The subsequent discussion ignores such easily-handled extensions to simplify the exposition.

Use of indicator functions provides an expression for $V_{(i)}$. One can characterize the set of values of c and h satisfying Equations (4.27) as

$$\left\{(c, h) \,\middle|\, I\left[h = \ell(\omega, y, z, \upsilon) \in (\overline{h}_{i-1}, \overline{h}_{i+1});\ b(c, h, W, Y, \xi) = 0\right] = 1\right\}, \tag{4.28}$$

where I denotes the indicator function defined by

$$I[\textit{conditions}] = \begin{cases} 1 & \text{if [}\textit{all conditions}\text{] are true,} \\ 0 & \text{if [}\textit{any condition}\text{] is false.} \end{cases}$$

The indicator function I in (4.28) depends on satisfaction of 2 conditions. Using I, a simple expression for the maximum utility attainable in state i is given by

$$V_{(i)} = V(\omega, y, z, \upsilon) * I\left[h = \ell(\omega, y, z, \upsilon) \in (\overline{h}_{i-1}, \overline{h}_{i+1});\ b(c, h, W, Y, \xi) = 0\right]. \tag{4.29}$$

For values of υ, η and ξ not yielding a solution in state i, $V_{(i)} = 0$. It is possible in this analysis for $V_{(i)} = 0$ for all values of admissible values of (υ, η, ξ) (i.e., $\Omega_i = \emptyset$). Throughout this discussion, we assume a utility function normalized so that $U(c, h, z, \upsilon) > 0$ for all admissible values of variables. So the event $V_{(i)} = 0$ always means that some state other than i has a higher assigned utility.

The most popular specifications of the budget function $b(c, h, W, Y)$ have derivatives b_h that depend on h but not on c. Examples include those specifications incorporating income or sales taxes given by (4.4). Under these circumstances, the first equation in (4.27) alone can be solved for h. Thus, $V_{(i)}$ simplifies to

$$V_{(i)} = V(\omega, y, z, v) * I\big[h = \ell(\omega, y, z, v) \in (\overline{h}_{i-1}, \overline{h}_{i+1})\big]. \tag{4.30}$$

This expression serves as the principal formulation used in the subsequent discussion.

The portion of a budget constraint selected by an individual depends on the level of utility assigned to this state. At kink points, utility takes the value

$$V_{(i)} = U(\overline{C}_i, \overline{h}_i, z, v), \tag{4.31}$$

where $\overline{C}_i$ and $\overline{h}_i$ designate the values of consumption and hours at the kink point associated with state i.

An individual occupies that portion of the budget constraint corresponding to state i if the assigned utility is highest for this state. According to (4.22), the subspace of (v, η, ξ) yielding this realization is the set Ω_i. Correspondingly, one can represent Ω_i as

$$\Omega_i = \big\{(v, \eta, \xi) \,\big|\, V_{(i)} > V_{(j)} \text{ for all } j \neq i\big\}. \tag{4.32}$$

Relationships (4.29) (or (4.30)) and (4.31) define $V_{(i)}$ depending on characteristics of the state. For expositional simplicity, without loss of generality, the subsequent discussion ignores equalities $V_{(i)} = V_{(j)}$ in defining the sets Ω_i since these events are zero probability events.

4.4.3. *Density functions for hours and wages*

The distribution of consumption and hours of work depends on where individuals locate on the budget constraints. The probability that an individual makes selections falling within the state i portion of the budget equals:

$$\begin{aligned} P(\delta_i = 1) &= P\big((v, \eta, \xi) \in \Omega_i\big) \\ &= \int \cdots \int_{\Omega_i} \varphi(v, \eta, \xi)\, dv\, d\eta\, d\xi \\ &\equiv \int_{\Omega_i} \varphi(v, \eta, \xi)\, dv\, d\eta\, d\xi. \end{aligned} \tag{4.33}$$

The notation $\int \cdots \int_{\Omega_i}$ denotes integration over the set Ω_i, which the third line of this equation expresses in the shorthand notation $\int_{\Omega_i}$. The joint distribution of the δ_i's takes the form

$$P(\delta_1, \ldots, \delta_m) = \prod_{i \in M} \big[P(\delta_i = 1)\big]^{\delta_i}$$

where the set M refers to the set of all possible states i that comprise the entire budget constraint. As noted previously, the events $\delta_i = 1$ may refer to either kinks or differentiable constraints.

When an optimum occurs at a kink point, the distribution of hours conditional on this event is

$$P(h = \bar{h}_i \mid \delta_i = 1) = 1. \tag{4.34}$$

This distribution is, of course, discrete.

On differentiable segments of the constraint, the distribution for hours is continuous. Performing a conventional change of variables yields the density

$$f(h, \eta, \xi) = \frac{dv^h}{dh}\varphi\big(v^h, \eta, \xi\big) = \frac{dv^h}{dh}\varphi\big(v^h(h, \omega, y, z), \eta, \xi\big) \tag{4.35}$$

where

$$v^h = v^h\big(h, \omega(h), y(h), z\big) = \ell^{-1}\big(h, \omega(h), y(h), z\big) \tag{4.36}$$

refers to the inverse of labor supply function (4.3), and the quantity

$$\frac{dv^h}{dh} = \left(\frac{\partial \ell}{\partial \omega}\frac{\partial \omega}{\partial h} - \frac{\partial \ell}{\partial y}\frac{\partial y}{\partial h}\right)\left(\frac{\partial \ell}{\partial v}\right)^{-1} \tag{4.37}$$

represents the Jacobian associated with this inverse. The terms $\frac{\partial \ell}{\partial \omega}$ and $\frac{\partial \ell}{\partial y}$ correspond to the economic concepts of substitution and income effects, and the quantity $\frac{\partial \ell}{\partial v}$ determines how unobserved components of preferences influence labor supply. (In applying the change-of-variables formula, Jacobians must be constructed to be uniquely signed for densities to be properly defined. This result follows here because the selection of budget-set partitions ensures a unique solution exists for c and h for each partition combined with the innocuous assumption that unobserved components enter preferences such that $\frac{\partial \ell}{\partial v} > 0$.) For the remaining terms in (4.37), differentiation of the budget constraint implies:

$$\frac{\partial \omega}{\partial h} = b_{hh} \tag{4.38}$$

and

$$\frac{\partial y}{\partial h} = -\frac{b_h}{b_c} - b_h - b_{hh}h \tag{4.39}$$

where the subscripts on the budget function b signify partial derivatives.[22] Assuming the popular tax form for the budget function given by (4.5), expressions (4.38) and (4.39)

[22] Derivation of the expression for $\frac{\partial y}{\partial h}$ follows from total differentiation of the relation (4.11) defining y which yields

$$b_c\left(b_h + b_{hh}h + \frac{\partial y}{\partial h}\right) + b_h = 0$$

and solving this equation.

simplify to

$$\frac{\partial \omega}{\partial h} = (1 - \tau')W \tag{4.40}$$

and

$$\frac{\partial y}{\partial h} = \tau'' W^2 h \tag{4.41}$$

where τ' and τ'' denote the marginal tax rate and its derivative. Division of the budget set into states ensures that inverse (4.36) and its Jacobian (4.37) exist in the space defined by each state.

The implied density of h conditional on δ_i is

$$f(h \mid \delta_i = 1) = \frac{\int_{\Phi_{i|h}} f(h, \eta, \xi)\, d\eta\, d\xi}{P(\delta_i = 1)} \quad \text{for } h \in \Theta_{i\cdot\, h} \tag{4.42}$$

where the set $\Theta_{i\cdot\, h} = (\overline{h}_{i-1}, \overline{h}_{i+1})$ designates the domain of h given occurrence of $\delta_i = 1$ and the notation $\int_{\Phi_{i|h}}$ denotes integration of (η, ξ) over the set

$$\Phi_{i|h} = \big\{(\eta, \xi) \,\big|\, I\big[h = \ell(\omega, y, z, v);\ (v, \eta, \xi) \in \Omega_i\big] = 1\big\}. \tag{4.43}$$

The set $\Phi_{i|h}$ treats h as fixed and, therefore, is a function of h.

Performing a further change of variables for wages yields the following joint density for hours and wages

$$f(h, W, \xi) = \frac{d\eta^w}{dW} f\big(h, \eta^w, \xi\big) = \frac{d\eta^w}{dW} f\big(h, \eta^w(W, Q), \xi\big) \tag{4.44}$$

where

$$\eta^w = \eta^w(W, Q) = W^{-1}(W, Q) \tag{4.45}$$

denotes the inverse of the wage function, and the quantity

$$\frac{d\eta^w}{dW} = \left[\frac{\partial W}{\partial \eta}\right]^{-1} \tag{4.46}$$

represents the Jacobian associated with this inverse. (For expositional convenience, and without loss of generality, this analysis assumes that a monotonically increasing relationship links W to η; so, (4.46) is positive.)

The density of h and W conditional on δ_i is

$$f(h, W \mid \delta_i = 1) = \frac{\int_{\Phi_{i|h,W}} f(h, W, \xi)\, d\xi}{P(\delta_i = 1)} \quad \text{for } (h, W) \in \Theta_{i\cdot\, h,W} \tag{4.47}$$

where the notation $\int_{\Phi_{i|h,W}}$ denotes integration of ξ over the set

$$\Phi_{i|h,W} = \big\{(\xi) \,\big|\, I\big[h = \ell(\omega, y, z, \upsilon);\ W = W(Q, \eta);\ (v, \eta, \xi) \in \Omega_i\big] = 1\big\}. \tag{4.48}$$

The set $\Phi_{i|h,W}$ is a function of h and W. One can express the set $\Theta_{i\cdot\, h,W}$ appearing in (4.47) as

$$\Theta_{i\cdot\, h,W} = \big\{(h, W) \,\big|\, \xi \in \Phi_{i|h,W}\big\},$$

which specifies the domain of h and W assuming occupancy of the state i part of the budget constraint.

If v is multi-dimensional as specified in labor supply function (4.13), then (4.47) becomes

$$f(h, W \mid \delta_i = 1) = \frac{\int_{\Phi_{i|h,W}} f(h, v_2, W, \xi)\, dv_2\, d\xi}{P(\delta_i = 1)} \quad \text{for } (h, W) \in \Theta_{i\cdot\, h,W} \tag{4.49}$$

where $f(h, v_2, W, \xi)$ has a form analogous to (4.44), and the notation $\int_{\Phi_{i|h,W}}$ now denotes integration of (v_2, ξ) over the set

$$\Phi_{i|h,W} = \big\{(v_2, \xi) \,\big|\, I\big[h = \ell(\omega, y, z, v);\ W = W(Q, \eta);\ (v, \eta, \xi) \in \Omega_i\big] = 1\big\}.$$

The set $\Phi_{i|h,W}$ still remains a function of h and W.

Finally, when an individual selects an optimum at a kink point and $h = \overline{h}$ is discrete, then the distribution of wages takes the form

$$f(\overline{h}, W \mid \delta_i = 1) = \frac{\int_{\Phi_{i|\overline{h},W}} f(v, W, \xi)\, dv\, d\xi}{P(\delta_i = 1)} \quad \text{for } W \in \Theta_{i\cdot\, W} \tag{4.50}$$

where the density $f(v, W, \xi)$ is specified analogous to (4.44), and the set $\Phi_{i|\overline{h},W}$ is a function of $\overline{h}$ and W defined by

$$\Phi_{i|\overline{h},W} = \big\{(v, \xi) \,\big|\, h = \overline{h}_i;\ W = W(Q, \eta);\ (v, \eta, \xi) \in \Omega_i\big\}. \tag{4.51}$$

The domain $\Theta_{i\cdot\, W}$ of W in (4.50) corresponds to that part of the overall range of W consistent with being at kink $\overline{h}_i$.

4.4.4. Likelihood functions for hours and wages

Appendix A presents the results required to develop a complete specification of the joint likelihood function for hours (h) and wages (W). Suppose the state $\delta_0 = 1$ refers to an individual choosing not to work; the states $\delta_i = 1$ for $i \in M_c$ designate those circumstances when the person works and selects optimums on differentiable segments of budget constraints; and the states $\delta_i = 1$ for $i \in M_d$ denote those events when an individual chooses hours located at a kink point. Hours (h) are continuously distributed for states in the set $i \in M_c$, and h is discretely distributed in the no-work state and for states in the set $i \in M_d$. Hours possess a combined continuous/discrete distribution. Knowledge of the value of h entirely determines the values of $\delta_0, \ldots, \delta_n$ where $n + 1$ designates the total number of states.

Formula (A.26) of Appendix A implies that the following specification delimits the joint likelihood function of (h, W):

$$\begin{aligned}\mathcal{L}(h, W) &= \mathcal{L}(h, W, \delta_0, \ldots, \delta_n) \\ &= \left[P\left((v, \eta, \xi) \in \Omega_0\right)\right]^{\delta_0} * \prod_{i \in M_c} \left[\int_{\Phi_{i|h,W}} f(h, v_2, W, \xi)\, dv_2\, d\xi\right]^{\delta_i} \\ &\quad * \prod_{i \in M_d} \left[\int_{\Phi_{i|\bar{h},W}} f(v, W, \xi)\, dv\, d\xi\right]^{\delta_i}. \end{aligned} \tag{4.52}$$

The first term of this expression delineates the probability of not working; the second term – comprised of the numerators of (4.49) – designates the densities of (h, W) unconditional on $\delta_i = 1$; and the third term – encompassing the numerators of (4.50) – demarcates the probability that $h = \bar{h}_i$ combined with the density of W unconditional on $\delta_i = 1$.

4.4.5. *Density functions accounting for measurement error*

With measurement error contaminating hours of work, h is no longer observed and one instead has data on measured hours H specified by relation (4.15). Without loss of generality, suppose (4.15) constitutes a monotonically increasing relationship that links H to the measurement error component ε. The joint density function (4.21) relates the distribution of ε to the distributions of the structural errors v, η, and ξ.

On differentiable segments of the budget constraint, the density function for true hours and wages is $f(h, W, \xi, \varepsilon)$ which has a form entirely analogous to (4.44). Performing a conventional change of variables yields the density

$$f(h, W, \xi, H) = \frac{\partial \varepsilon^H}{\partial H} f\left(h, W, \xi, \varepsilon^H\right) = \frac{\partial \varepsilon^H}{\partial H} f\left(h, W, \xi, \varepsilon^H(H, h)\right) \tag{4.53}$$

where

$$\varepsilon^H = \varepsilon^H(H, h) = H^{-1}(H, h) \tag{4.54}$$

refers to the inverse of measurement error function (4.15), and the quantity

$$\frac{\partial \varepsilon^H}{\partial H} = \left[\frac{\partial H}{\partial h}\right]^{-1} \tag{4.55}$$

designates the Jacobian associated with this inverse. The corresponding density of H and W conditional on $\delta_i = 1$ is

$$\begin{aligned} f(H, W \mid \delta_i = 1) &= \frac{\int_{\Theta_{i\cdot\, h}} \int_{\Phi_{i|h,\widetilde{W}}} f(h, W, \xi, H)\, d\xi\, dh}{P(\delta_i = 1)} \\ &\quad \text{for } (H, W) \in \Theta_{i\cdot\, H,W} \end{aligned} \tag{4.56}$$

where $\int_{\Theta_{i\cdot\, h}}$ denotes integration over the set $\Theta_{i\cdot\, h}$ which corresponds to the domain of h conditional on $\delta_i = 1$.

When wages are also measured with error through mismeasurement of hours as characterized by relation (4.16), then (4.53) is replaced by

$$f(h, \widetilde{W}, \xi, H) = f\big(h, \widetilde{W}^{-1}(\widetilde{W}, h, H), \xi, H\big) \tag{4.57}$$

where

$$W = \widetilde{W}^{-1}\big(\widetilde{W}, h, \varepsilon^H(H, h)\big) \equiv \widetilde{W}^{-1}(\widetilde{W}, h, H)$$

refers to the inverse of measurement error function (4.16). The corresponding density of H and $\widetilde{W}$ conditional on $\delta_i = 1$ becomes

$$f(H, \widetilde{W} \mid \delta_i = 1) = \frac{\int_{\Theta_{i\cdot\, h}} \int_{\Phi_{i|h,W}} f(h, \widetilde{W}, \xi, H)\, d\xi\, dh}{P(\delta_i = 1)}$$
$$\text{for } (H, \widetilde{W}) \in \Theta_{i\cdot\, H,\widetilde{W}}. \tag{4.58}$$

No change of variables occurs in deriving this expression since $\widetilde{W}$ is fully known given values for h, W and H.

A similar situation applies to incorporating measurement error when an individual selects an optimum at a kink point of the budget set. Conditional on realization of the state $\delta_i = 1$, the value of ε is known since one sees H and $h = \overline{h}_i$ with probability one. Defining the $f(\nu, W, \xi, \varepsilon)$ as the generalization of the joint density function appearing in (4.50) incorporating measurement error, then substitution of the inverse functions $\widetilde{W}^{-1}(\widetilde{W}, \overline{h}_i, H)$ and $\varepsilon^H(H, \overline{h}_i)$ introduced above into this joint density yields

$$f(\nu, \widetilde{W}, \xi, H) = f\big(\nu, \widetilde{W}^{-1}(\widetilde{W}, \overline{h}_i, H), \xi, \varepsilon^H(H, \overline{h}_i)\big). \tag{4.59}$$

Following the steps above, one can readily verify that the density of $(H, \widetilde{W})$ conditional on $\delta_i = 1$ takes the form

$$f(H, \widetilde{W} \mid \delta_i = 1) = \frac{\int_{\Phi_{i|\overline{h},\widetilde{W}}} f(\nu, \widetilde{W}, \xi, H)\, d\nu\, d\xi}{P(\delta_i = 1)} \quad \text{for } \widetilde{W} \in \Theta_{i\cdot\, \widetilde{W}}. \tag{4.60}$$

Clearly, both specifications (4.59) and (4.60) depend directly on $\overline{h}_i$, but as in representation of other specifications, the only arguments included in the function are those variables that are random in the state; $\overline{h}_i$ is fixed and known given $\delta_i = 1$.

4.4.6. Likelihood functions for measured hours and wages

Formulating the likelihood function for $(H, \widetilde{W})$ is complicated by the fact that a researcher does not observe precisely which portion of the budget constraint an individual selects since this decision reveals h and this quantity is unknown. Thus, when a person works, one cannot distinguish which individual state i occurs. On the other hand,

a researcher does observe when a person does not work. Expressed in terms of the endogenous dummy variables δ_i, these circumstances imply that the data reveal the event $\delta_0 = 1$ but not the individual events $\delta_i = 1$ for $i \in M = M_c \cup M_d$. Instead, one merely observes whether $\overline{\delta}_1 \equiv \sum_{i \in M} \delta_i = 1$ or $\overline{\delta}_1 = 0$.

Appealing to formula (A.34) of Appendix A, the following specification represents the joint likelihood function of (h, W):

$$\begin{aligned}\mathcal{L}(H, \widetilde{W}) = {} & \left[P\left((v, \eta, \xi) \in \Omega_0\right)\right]^{\overline{\delta}_0} \\ & * \left[\sum_{i \in M_t} \int_{\Theta_i} \left[\int_{\Phi_{i|h,W}} f(h, v_2, \widetilde{W}, \xi, H)\, dv_2\, d\xi\right] dh \right. \\ & \left. + \sum_{i \in M_t} \int_{\Phi_{i|\overline{h},W}} f(v, \widetilde{W}, \xi, H)\, dv\, d\xi\right]^{\overline{\delta}_t}. \end{aligned} \tag{4.61}$$

The first term of this expression delineates the probability of not working; and the second term designates the density of $(H, \widetilde{W})$ unconditional on $\overline{\delta}_1$. Accordingly, both H and $\widetilde{W}$ are continuously distributed throughout the range on $H > 0$.

4.5. Maximum likelihood: Convex differentiable constraints with full participation

Developing specifications for likelihood functions when budget sets are convex and have differentiable boundaries is straightforward, especially assuming labor force participation is not a factor for the population under investigation. The following discussion presents two examples of such specifications to illustrate elementary versions of the general formulas presented above.

4.5.1. Specifications for linear parameterizations of labor supply

Derivation of likelihood functions assuming a linear specification for hours of work when (4.5) describes the budget constraint – wherein tax payments depend only on a single taxable income quantity – follows directly from the previous results. Assuming no measurement error (i.e., $H = h$), a change in variables from the heterogeneity error v to actual hours h yields the likelihood function for h:

$$f_h(h) = \frac{dv}{dh}\varphi_v(h - y_v - z\gamma - \alpha\omega - \beta y) \tag{4.62}$$

where $\varphi_v(v)$ denotes the density of the heterogeneity component v, and the Jacobian term is

$$\frac{dv}{dh} = 1 + (\alpha - \beta h)W^2 \frac{\partial \tau'}{\partial I}. \tag{4.63}$$

This Jacobian term is restricted to be nonnegative over the admissible range. Maximizing (4.62) yields maximum likelihood estimates for the parameters of the labor supply

function, ℓ, which provide the information needed to infer the work disincentive effects of taxation.

If hours are indeed contaminated by additive measurement error, then the likelihood function for observed hours $H = h + \varepsilon$ is given by

$$f_H(H) = \int_0^{\text{max hours}} \varphi_\varepsilon(H - h)\varphi_h(h)\,dh \tag{4.64}$$

where $\varphi_\varepsilon(\varepsilon)$ denotes the density of the heterogeneity component ε. This expression resembles relation (4.62) except that integration occurs over hours to account for the existence of reporting error, and H replaces actual hours h in the Jacobian term in (4.63).

4.5.2. *Specifications for multiplicative measurement error*

Now consider maximum likelihood estimation of the semilog specification of labor supply. Suppose the heterogeneity error component ν in the structural labor supply equation and the disturbance ε in the measurement error equation for hours of work possess the joint distribution $\varphi_{\nu\varepsilon}(\nu, \varepsilon)$, where $\varphi_{\nu\varepsilon}$ designates a density function. For the moment, suppose (ν, ε) are distributed independently of the gross wage and other income. Using relations (4.35) and (4.44) to perform a standard change in variables from the errors ν and ε to the variables h and H produces the likelihood function needed to compute maximum likelihood estimates. The transformation from (ν, ε) to (h, H) is monotonic for a wide range of functional forms for ℓ as long as the underlying preferences satisfy quasiconcavity and budget sets are convex.

Without measurement error, the likelihood function for hours of work, h, takes the form

$$f_h(h) = \frac{d\nu}{dh}\varphi_\nu\big(h - y - z\gamma - \alpha \ln W - \alpha \ln(1 - \tau') - \beta y\big) \tag{4.65}$$

where φ_ν is the marginal density for ν, and the Jacobian term is

$$\frac{d\nu}{dh} = 1 + \left(\left(\frac{\alpha}{W(1 - \tau')} - \beta h\right)W^2\frac{\partial \tau'}{\partial I}\right) \tag{4.66}$$

which is required to be nonnegative. In these expressions, the derivative τ' is evaluated at $I = Wh + Y - \tau(Wh + Y)$.

With multiplicative measurement error, the likelihood function for observed hours H becomes

$$\begin{aligned}\mathcal{L} = \int_0^{\text{max wage}} \int_0^{\text{max hours}} &\frac{d\nu}{dh} \\ &\times f_{\nu\varepsilon w}(h - y - z\gamma - \alpha \ln \omega - \beta y, \ln H - \ln h, W)\,dh\,dW\end{aligned} \tag{4.67}$$

where integration occurs over the hourly wage, which is unobserved, using the joint density $f_{\nu\varepsilon w}(\nu, \varepsilon, W)$. The nonnegativity of the Jacobian term clearly places restrictions on the behavioral parameters and we discuss these restrictions further below.

4.6. Maximum likelihood: Convex piecewise-linear constraints with full participation

The majority of empirical labor supply studies incorporating taxes treat the tax schedule as a series of brackets implying a piecewise-linear budget set. With such a tax function, the familiar change-in-variables techniques implemented in conventional maximum likelihood do not apply due to the nonexistence of the Jacobian over measurable segments of the sample space arising from nondifferentiability of functional relationships characterizing hours-of-work choices. Moreover, a piecewise-linear budget set creates endogenous variables (hours and after-tax wages) that are both discrete and continuous in character. Section 4.4 covers specifications for likelihood functions for such endogenous variables.

4.6.1. Characterization of labor supply with piecewise-linear constraints

To illustrate the derivation of an estimable labor supply model using the piecewise-linear approach assuming the linear structural specification for hours of work, consider the simple case of a budget set with only three segments as presented in Figure 4.1. The preceding discussion defines the variables y_j, ω_j and $\overline{h}_j$ appearing in this figure. To locate the kinks and slopes of the budget constraint for an individual, a researcher must know the individual's level of nonlabor income, gross wage rate, hours of work, and the structure of the tax system. The hours of work at which kinks occur are given by $\overline{h}_j = (I_j - Y + D)/W$, where Y and D, respectively, represent taxable nonlabor income and deductions, and I_j is the maximum taxable income for segment j. The slope of each segment is given by the marginal wage rate for that segment: $\omega_j = W(1 - t_j)$, where j denotes the segment, t_j signifies the marginal tax rate for that segment, and W is the gross wage rate per hour. Finally, the nonlabor income at zero hours of work – the intercept of the budget line – is $y_1 = Y - \tau(Y - D)$, where $\tau(\cdot)$ is the tax function evaluated at the individual's taxable income at zero earnings. Given this intercept value, virtual incomes or the intercepts associated with successive budget segments are computed by repeated application of the formula: $y_j = y_{j-1} + (\omega_{j-1} - \omega_j)\overline{h}_{j-1}$.

Given a convex budget constraint, an individual's optimization problem amounts to maximizing $U(c, h)$ subject to

$$c = \begin{cases} y_1 & \text{if } h = 0, \\ \omega_1 h + y_1 & \text{if } H_0 < h \leqslant \overline{h}_1, \\ \omega_2 h + y_2 & \text{if } \overline{h}_1 < h \leqslant H_1, \\ \omega_3 h + y_3 & \text{if } H_1 < h \leqslant \overline{h}_3, \\ \omega_3 \overline{h}_3 + y_3 & \text{if } h = \overline{h}_3. \end{cases} \tag{4.68}$$

The solution of this maximization problem decomposes into two steps. First, determine the choice of h conditional on locating on a particular segment or a kink. This step

yields the solution:

$$h = \begin{cases} 0 & \text{if } h = 0 \quad \text{(lower limit)}, \\ \ell(\omega_1, y_1, v) & \text{if } 0 < h < \overline{h}_1 \quad \text{(segment 1)}, \\ \overline{h}_1 & \text{if } h = \overline{h}_1 \quad \text{(kink 1)}, \\ \ell(\omega_2, y_2, v) & \text{if } \overline{h}_1 < h < H_1 \quad \text{(segment 2)}, \\ H_1 & \text{if } h = H_1 \quad \text{(kink 2)}, \\ \ell(\omega_3, y_3, v) & \text{if } H_1 < h < \overline{h}_3 \quad \text{(segment 3)}, \\ \overline{h}_3 & \text{if } h = \overline{h}_3 \quad \text{(kink 3 = upper limit)}. \end{cases} \tag{4.69}$$

Second, determine the segment or the kink on which the person locates. The following relations characterize this solution: choose

$$\begin{aligned} &0 && \text{if } \ell(\omega_1, y_1, v) \leqslant 0, \\ &\text{(Segment 1)} && \text{if } H_0 < \ell(\omega_1, y_1, v) < \overline{h}_1, \\ &\text{(Kink 1)} && \text{if } \ell(\omega_2, y_2, v) \leqslant \overline{h}_1 < \ell(\omega_1, y_1, v), \\ &\text{(Segment 2)} && \text{if } \overline{h}_1 < \ell(\omega_2, y_2, v) < H_1, \\ &\text{(Kink 2)} && \text{if } \ell(\omega_3, y_3, v) \leqslant \overline{h}_2 < \ell(\omega_2, y_2, v), \\ &\text{(Segment 3)} && \text{if } H_1 < \ell(\omega_3, y_3, v) < \overline{h}_3, \\ &\text{(Kink 3)} && \text{if } \ell(\omega_3, y_3, v) \geqslant \overline{h}_3. \end{aligned} \tag{4.70}$$

Combined, these two steps imply the values of h and C that represent the utility-maximizing solution for labor supply and consumption.

All studies implementing the piecewise-linear approach assume the existence of measurement error in hours of work. With the linear measurement error model observed hours $H = h + \varepsilon$. As long as the measurement error component ε is continuously distributed, so is H. In contrast to information on h, knowledge of H suffices neither to allocate individuals to the correct branch of the budget constraint nor to identify the marginal tax rate faced by individuals, other than at zero hours of work. The state of the world an individual occupies can no longer be directly observed, and one confronts a discrete-data version of an errors-in-variables problem. The interpretation of measurement error maintained in this analysis is that ε represents reporting error that contaminates the observation on h for persons who work.[23]

[23] Note that expected hours of work, in this convex piecewise-linear case, is additive in each hours choice weighted by the probability of each segment or kink, each term in this sum being at most a function of two marginal wages and two virtual incomes. Blomquist and Newey (2002) exploit this observation to develop a semiparametric estimator for hours of work imposing the additivity through a series estimator.

With measurement error, the linear specification of labor supply with $\hat{h}_j \equiv \mu + \alpha\omega_j + \beta y_j + Z\gamma$ implies the following stochastic specification:

$$H = \begin{cases} \hat{h}_1 + \nu + \varepsilon & \text{if } 0 < \hat{h}_1 + \nu \leqslant \overline{h}_1 \quad \text{(segment 1)}, \\ \overline{h}_1 + \varepsilon & \text{if } \hat{h}_2 + \nu < \overline{h}_1 < \hat{h}_1 + \nu \quad \text{(kink 1)}, \\ \hat{h}_2 + \nu + \varepsilon & \text{if } \overline{h}_1 < \hat{h}_2 + \nu \leqslant H_1 \quad \text{(segment 2)}, \\ H_1 + \varepsilon & \text{if } \hat{h}_3 + \nu < \overline{h}_2 < \hat{h}_2 + \nu \quad \text{(kink 2)}, \\ \hat{h}_3 + \nu + \varepsilon & \text{if } H_1 < \hat{h}_3 + \nu \leqslant \overline{h}_3 \quad \text{(segment 3)}, \\ \overline{h}_3 + \varepsilon & \text{if } \hat{h}_3 + \nu \geqslant \overline{h}_3 \quad \text{(upper limit)}. \end{cases} \tag{4.71}$$

This represents a sophisticated variant of an econometric model that combines discrete and continuous choice elements.

4.6.2. Likelihood function with measurement error when all work

The log-likelihood function for this model is given by $\sum_i \ln f_H(H)$, where i indexes observations. Defining $\underline{v}_j = \overline{h}_{j-1} - \hat{h}_j$ and $\overline{v}_j = \overline{h}_j - \hat{h}_j$, the components $f_H(H)$ are given by

$$\begin{aligned} f_H(H) = &\sum_{j=1}^{3} \int_{\underline{v}_j}^{\overline{v}_j} \varphi_2(H - \hat{h}_j, \nu)\, d\nu \quad \text{(segments 1, 2, 3)} \\ &+ \sum_{j=1}^{2} \int_{\overline{v}_j}^{\underline{v}_{j+1}} \varphi_1(H - \overline{h}_j, \nu)\, d\nu \quad \text{(kinks 1, 2)} \\ &+ \int_{\overline{v}_3}^{\infty} \varphi_1(H - \overline{h}_3, \nu)\, d\nu \quad \text{(upper limit)} \end{aligned} \tag{4.72}$$

where $\varphi_1(\cdot,\cdot)$ and $\varphi_2(\cdot,\cdot)$ are the bivariate density functions of (ε, ν) and $(\varepsilon + \nu, \nu)$, respectively. Maximizing the log-likelihood function produces estimates of the coefficients of the labor supply function ℓ. These estimates provide the information used to infer both substitution and income responses, which in turn provide the basis for calculating the work disincentive effects of income taxation.

4.6.3. Shortcomings of conventional piecewise-linear analyses

The piecewise-linear approach for estimating the work disincentive effects of taxes offers both advantages and disadvantages relative to other methods. Concerning the attractive features of this approach, piecewise-linear analyses recognize that institutional features of tax systems induce budget sets with linear segments and kinks. This is important if one believes that a smooth tax function does not provide a reasonably accurate description of the tax schedule. The piecewise-linear approach admits randomness in hours of work arising from both measurement error and variation in individual

preferences and it explicitly accounts for endogeneity of the marginal tax rate in estimation, but so do the instrumental variable and differentiable likelihood methods discussed above. As we will see below, the piecewise-linear approach more readily incorporates fixed costs of holding a job, regressive features of the tax code, and multiple program participation than other procedures due to the discrete-continuous character of hours-of-work choices induced in these environments. These features of the piecewise-linear method make it a vital approach in empirical analyses of labor supply.

On the other hand, the following shortcomings of the piecewise-linear procedure raise serious doubts about the reliability of its estimates of work disincentive effects. First, the piecewise-linear methodology assumes that both the econometrician and each individual in the sample have perfect knowledge of the *entire* budget constraint that is relevant for the worker in question. Errors are permitted neither in perceptions nor in measuring budget constraints. Taken literally, this means that: all income and wage variables used to compute each sample member's taxes are observed perfectly by the econometrician; individuals making labor supply choices know these variables exactly prior to deciding on hours of work; each individual and the econometrician know when the taxpayer will itemize deductions and the amount of these itemizations; and each taxpayer's understanding of the tax system is equivalent to that of the econometrician (e.g., the operation of such features as earned-income credits). Clearly, given virtual certainty that most of these assumptions are violated in empirical analyses of labor supply, the estimates produced by methods relying on these assumptions must be interpreted very cautiously. The differentiable likelihood methods rely on the same assumptions. The instrumental variable methods do not, so they are likely to be more robust.

Second, measurement error plays an artificial role in econometric models based on the piecewise-linear approach. Its presence is needed to avoid implausible predictions of the model. The statistical framework induced by the piecewise-linear approach implies that bunching in hours of work should occur at kink points if hours precisely measure h. However, for the vast majority of data sources currently used in the literature, only a trivial number of individuals, if indeed any at all, report hours of work at interior kink points. Unless one presumes that the data on hours do not directly represent h, such evidence provides the basis for immediately rejecting the distributional implications of the above specifications. Considering, for example, the labor-supply characterization proposed in Equation (4.69), almost any test of the distributional assumptions implied by this specification would be readily rejected because observed hours would take the values H_0, $\overline{h}_1$, $\overline{h}_2$ and $\overline{h}_3$ with only a trivial or zero probability. Instead, observed hours essentially look as if they are distributed according to a continuous distribution. When a continuously-distributed measurement error ε is added to the model, observed hours H are continuously distributed. This provides an essential reason for introducing measurement error in the data, for without it, the piecewise-linear structure provides a framework that is grossly inconsistent with the data. Of course, several sound reasons exist for admitting measurement error in a labor supply model, including the widespread suspicion that reporting error contaminates data

on hours of work. However, measurement error in hours of work implies measurement error in wages, since they are typically computed as average hourly earnings. Current applications of the piecewise-linear analysis mistakenly ignore this by assuming perfectly measured budget constraints.[24] The unnatural role played by measurement error raises questions about the credibility of findings derived from the piecewise-linear approach. In contrast to the piecewise-linear approach, it is not essential to introduce measurement error in either the differentiable likelihood or the instrumental variable approach because hours in the distribution of h are continuous without measurement error.

Third, existing research implementing the piecewise-linear methodology relies on very strong exogeneity assumptions. Other than hours of work, all variables involved in the calculation of taxes are presumed to be exogenous determinants of labor supply behavior, both from a statistical and from an economic perspective. These variables include gross wages, the various components of nonlabor income, and deductions. In light of the evidence supporting the view that wages and income are endogenous variables in labor supply analyses, particularly in the case of wages, suspicions arise regarding the dependability of estimated substitution and income effects based on procedures that ignore such possibilities. Most of the exogeneity assumptions are also maintained in the differentiable likelihood approach, but are easily relaxed when applying instrumental variable procedures (given the availability of a sufficient number of other instrumental variables).

Fourth, some concerns about the reliability of estimates produced by the piecewise-linear approach ensue due to the static behavioral framework maintained in the formulation of empirical relations. Piecewise-linear studies invariably rely on the textbook one-period model of labor supply as a description of hours-of-work choices, and impose it to estimate parameters. Existing implementations of the differentiable likelihood approach suffer from the same problem. Everyone acknowledges that individuals are not simply myopic optimizers; they transfer income across periods to achieve consumption plans that are infeasible without savings. A serious question arises concerning the relevance of such considerations in estimating substitution and income effects used to predict responses to tax policy.

4.7. Maximum likelihood estimation imposes restrictions on behavioral responses

The implementation of maximum likelihood procedures imposes interesting and important restrictions on behavioral parameters in the presence of nonlinear budget constraints. These restrictions come about in defining the statistical model to be coherent, requiring probabilities to fall in the [0, 1] interval and densities to be nonnegative.

[24] It is possible to argue that this error does not result in measurement error in the hourly wage, if the measurement error is interpreted as an "optimization" error.

4.7.1. Restrictions imposed under piecewise-linear constraints

The econometric model produced by the piecewise-linear formulation given by (4.72) implicitly imposes parametric restrictions that constrain the signs of estimated substitution and income effects. As developed in MaCurdy, Green and Paarsch (1990), particular inequality restrictions must hold in the application of estimation procedures with piecewise-linear budget constraints for likelihood functions to be defined (i.e., to ensure that the components of these functions are nonnegative). More specifically, in applications of such procedures, the Slutsky condition must be locally satisfied at all interior kink points of budget sets that represent feasible options for any individual in the sample such that the compensated substitution effect must be positive. For the linear specification of the labor supply function considered in the preceding discussion, the specific inequality constraints imposed are

$$\alpha - \beta \overline{h}_{jk} \geqslant 0, \quad \forall j, k, \tag{4.73}$$

where the quantities $\overline{h}_{jk}$ represent the hours-of-work values that correspond to interior kink points j on a sample member k's budget set. Because many values of $\overline{h}_{jk}$ exist in most analyses of piecewise-linear constraints, fulfillment of relations (4.73) essentially requires global satisfaction of the Slutsky condition by the labor supply function. Such a requirement, in essence, globally dictates that the uncompensated substitution effect of a wage change on hours of work must be positive for the labor supply specification considered in the preceding discussion, and the income effect for hours of work must be negative. The imposition of these restrictions, especially for men, is highly suspect given the available evidence from other studies. These restrictions carry over to more general labor supply functions.

4.7.2. Restrictions imposed under differentiable constraints

Maximum likelihood estimation with differentiable constraints induces comparable restrictions. Consider, for example, likelihood function (4.62). For this specification to be a properly-defined likelihood function, the Jacobian $\frac{dv}{dh}$ must be nonnegative. Violation of this condition implies that the density function for h is negative, which obviously cannot occur. Nonnegativity of $\frac{dv}{dh}$ translates into the property

$$\frac{\partial \ell}{\partial \omega} - \frac{\partial \ell}{\partial y} h \geqslant -\left(\frac{\partial \tau}{\partial I} W^2\right)^{-1} \leqslant 0, \tag{4.74}$$

where ℓ refers to the labor supply function. The left-hand side of this inequality is the Slutsky term. This inequality result does not require compensated substitution effects to be positive as quasiconcave preferences mandate, only that these effects cannot become too negative.

Maximum likelihood procedures yield nonsensical results unless Equation (4.74) holds. Without measurement error, estimated parameter values cannot imply a violation

of Equation (4.74) at any of the data combinations $(h, \omega(h), y(h))$ actually observed in the sample. If a violation occurs, then the evaluation of (4.62) for the observation associated with this combination would result in a nonpositive value which causes the overall log likelihood function to approach minus infinity which clearly cannot represent a maximum.

With measurement error, maximum likelihood estimation applied to function (4.64) ensures that a weighted average of density functions appearing in (4.64) holds, with weighting occurring over all combinations of hours, marginal wages, and virtual income lying in the feasible range of the budget constraint of any individual included in the sample. Since maximum likelihood procedures assume the validity of such restrictions when calculating estimates of the coefficients of ℓ, the resulting estimated labor supply function can be expected to exhibit compensated substitution effects that obey inequality (4.74) over a very wide range of hours, wages, and incomes.[25]

4.8. Maximum likelihood: Accounting for participation and missing wages

As mentioned in previous sections, some applications of the piecewise-linear approach incorporate fixed costs to working – costs such as transportation that must be paid for any amount of work but which may vary across individuals. This significantly complicates the analysis because the optimized level of work under the budget constraint while working may not represent the optimal choice overall; one must explicitly consider the option of not working and thus avoiding the fixed cost. For any level of fixed costs, a minimum number of hours worked is implied creating an attainable range in the observable hours of work distribution; individuals will not work unless the gain is large enough to overcome the fixed costs. In essence, these complications arise because the budget constraint is not convex, invalidating simple maximization procedures.

4.8.1. Fixed costs of working

If an individual must pay fixed monetary costs, F, to work, then nonlabor income, Y, in the above budget constraints is replaced by

$$\begin{aligned} &Y - F \quad \text{if } h > 0, \\ &Y \qquad \text{if } h = 0. \end{aligned} \tag{4.75}$$

F is partially unobservable and, thus, modeled as a stochastic element, varying across individuals. Hence, we see that the budget constraint discontinuously jumps down by F when the individual chooses to work.

[25] It is, of course, computationally feasible to use (4.64) in estimation and not require f_h to be defined over the entire range of its support. Computationally one merely requires f_h to be nonnegative over a sufficiently large region to ensure (4.64) > 0. Of course, not requiring $f_h \geqslant 0$ over its relevant range produces a nonsensical statistical model.

To solve for the optimum when faced with this budget constraint, two regimes must explicitly be considered: working and not working. Estimation proceeds by finding the maximum utility under each regime and then comparing these to determine which option is chosen. In neither regime, the utility function $U(c, h, \nu)$ – where we explicitly note the unobserved component, ν – is maximized subject to optimization problem (4.1) with (4.4) modified by (4.75).

In the no-work regime, the solution is simple. We know h is 0, so utility is given by $U(Y - \tau(Y - D), 0, \nu)$.

The solution in the work regime closely follows the solution presented in Section 4.6. Again utilizing the labor supply function, $\ell(\omega, y, \nu)$ yields the solution for h given in (4.69), where the virtual income y now subtracts fixed costs F. However, to compute maximum utility in this regime requires associating a utility level with each possible hours choice. Utility along any segment, j, is given by the indirect utility function, $V(\omega_j, y_j, \nu)$. At kinks, the direct utility function must be used, so the utility at kink j is given by $U(\omega_j \overline{h}_j + y_j, \overline{h}_j, \nu)$. Hence, utilizing exactly the same solution procedure exploited in Section 4.6, we can define maximized utility when working, V^*:

$$V^*(w, y, \nu) = \begin{cases} -\infty, & \ell_1 \leqslant 0, \\ V(\omega_1, y_1, \nu), & 0 < \ell_1 < \overline{h}_1, \\ U(\omega_1 \overline{h}_1 + y_1, \overline{h}_1, \nu), & \ell_2 < \overline{h}_1 \leqslant \ell_1, \\ V(\omega_2, y_2, \nu), & \overline{h}_1 < \ell_2 < H_1, \\ U(\omega_2 H_1 + y_2, H_1, \nu) & \ell_3 < H_1 \leqslant \ell_2, \\ V(\omega_3, y_3, \nu), & H_1 < \ell_3 < \overline{h}_3, \\ U(\omega_3 \overline{h_m} + y_3, \overline{h_m}, \nu), & \ell_3 \geqslant \overline{h_m}, \end{cases} \tag{4.76}$$

where

$$\ell_j \equiv \ell(\omega_j, y_j, \nu) \equiv \frac{V_\omega(\omega_j, y_j, \nu)}{V_y(\omega_j, y_j, \nu)} \tag{4.77}$$

with V_ω and V_y denoting the partial derivatives of V; relation (4.77) is, of course, Roy's identity defining the labor supply function, ℓ, evaluated at wage and income levels ω_j and y_j. The use of $-\infty$ for $h = 0$ simply indicates that $h = 0$ is not included in this regime and, thus, selecting it indicates that the no-work regime is preferred. Given functional forms for V and U, finding V^* is straightforward.

Given maximized utility under each regime, the final step in the solution is to compare the two regimes. An individual chooses to work at the hours specified by the solution in (4.69) if

$$V^*(\omega, y, \nu) \geqslant U\big(Y - \tau(Y - D), 0, \nu\big) \tag{4.78}$$

and chooses not to work otherwise. For any level of ν, treating Equation (4.78) as an equality implies a critical level of fixed costs, $F^*(\nu)$ above which the individual will choose not to work; F enters this relation through the virtual income variable y. Because desired hours of work increase with ν, this critical value will generally be increasing in

ν – greater propensity to work implies that higher fixed costs are required to prefer the no-work option. If restrictions are placed on the support of F, such as $F > \underline{F}$, there will be values of ν low enough to rule out the work regime, thus implying a hole at the low end of the h distribution.

As a final step before deriving the likelihood function, note that in the no-work regime, gross wage, W, is not observed and, thus, the budget constraint cannot be derived. Hence, W must be endogenized. Such a step amounts to modeling the offered gross wage rate as being generated by a variant of Equation (2.24) which presumes that W is randomly distributed across the population depending on measured characteristics $Q = (x, q)$ and unobservable components η i.e., $W = W^*(Q) + \eta$.

4.8.2. *Likelihood function incorporating fixed costs*

To derive the likelihood function, first consider the likelihood contribution of an individual who does not work. We assume this no-work decision can be observed, so there is no measurement error. In the no-work case, one of two situations applies: (i) fixed costs are sufficiently high with $F > F^* \equiv F^*(\nu, \eta)$ for any given ν and η, or (ii) if this fixed-cost threshold falls below the lowest admissible value for F (i.e. $F^* \leqslant \underline{F}$), then desired hours are sufficiently low with $\nu < \nu^* \equiv \nu^*(\eta)$ for any η.[26] The probability of this event is

$$\mathcal{L}_0 = \int_{-\infty}^{\infty} \int_{-\infty}^{\nu^*} \int_{F^*}^{\infty} \varphi_{\nu\eta F}(\nu, \eta, F)\, dF\, d\eta\, d\nu \tag{4.79}$$

where $\varphi_{\nu\eta F}$ is the joint density of (ν, η, F).

For the work regime, the likelihood contribution looks very much like that derived in specification (4.72), as we continue to assume the linear hours of work function and the form of measurement error assumed there. The only changes are the addition of terms for δ and F (accounting for the fact that $F < F^*(\nu)$) and the removal of the term for the lower limit which is no longer part of that regime and is now perfectly observable. Using φ_1 and φ_2 to denote the distribution of $(\varepsilon, \nu, \eta, F)$ and $(\varepsilon + \nu, \nu, \eta, F)$ yields:

$$\begin{aligned}
\mathcal{L}_1 = {} & \sum_{j=1}^{3} \int_{\underline{\nu}_j}^{\overline{\nu}_j} \int_{0}^{F^*} \varphi_2\big(H - \hat{h}_j, \nu, W - W^*(Q), F\big)\, dF\, d\nu \\
& + \sum_{j=1}^{2} \int_{\overline{\nu}_j}^{\underline{\nu}_{j+1}} \int_{0}^{F^*} \varphi_1\big(H - \overline{h}_j, \nu, W - W^*(Q), F\big)\, dF\, d\nu \\
& + \int_{\overline{\nu}_3}^{\infty} \int_{0}^{F^*} \varphi_1\big(H - \overline{h}_3, \nu, W - W^*(Q), F\big)\, dF\, d\nu
\end{aligned} \tag{4.80}$$

[26] The critical value v^* solves relation (4.78) treated as an equality with virtual income y evaluated at $\underline{F}$.

where

$$\underline{v}_j \text{ solves the equation } \ell(\omega_j, y_j, \underline{v}_j) = \overline{h}_{j-1},$$
$$\overline{v}_j \text{ solves the equation } \ell(\omega_j, y_j, \overline{v}_j) = \overline{h}_j. \tag{4.81}$$

All variables appearing in these expressions are defined as in Section 4.6.

The likelihood function for an individual is given by

$$\mathcal{L} = (\mathcal{L}_1)^{\delta_E}(\mathcal{L}_0)^{1-\delta_E} \tag{4.82}$$

where $\delta_E = 1$ if the individual works and $\delta_E = 0$ otherwise. Estimation proceeds by maximizing the sum of log likelihoods across individuals, as always. This is quite complex in this case, requiring knowledge of both the direct utility U and the indirect utility V, and also requiring comparisons across regimes for all individuals and all parameter values.

4.9. Welfare participation: Maximum likelihood with nonconvex constraints

A common source of nonlinearity in budget constraints involves participation in welfare programs. To illustrate this situation, consider the simplest case in which the only taxes faced by an individual result from benefit reduction on a single welfare program. Figure 4.3 presents this scenario. Under most welfare programs, individuals face very high effective tax rates when they initially work due to large reductions in their benefits occurring when earnings increase. Once benefits reach 0, the tax rate drops to a lower level, creating a nonconvex kink in the budget constraint. This nonconvexity invalidates the simple procedures exploited in Section 4.6 implemented to divide sample spaces into locations on budget sets.

4.9.1. Simple nonconvex constraints with no welfare stigma

Following the picture portrayed in Figure 4.3, an individual maximizes $U(c, h, v)$ subject to the budget constraint

$$c = Wh + Y + b\big(I(h)\big), \tag{4.83}$$

where benefits are given by the simple benefit schedule:

$$b\big(I(h)\big) = \begin{cases} G - \rho Wh & \text{if } G - \rho Wh > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.84}$$

G gives the guarantee amount which is reduced at the benefit reduction rate ρ as the earnings, Wh, increase. This implies a kink point at $H_1 = G/\rho W$ where benefits reach 0 and, thus, the marginal wage rises to W. So, the individual faces two segments: seg-

ment 1 has $h < \overline{h}_1$ with net wage $\omega_1 = (1-\rho)W$ and virtual income $y_1 = Y + G$; and segment 2 has $h > \overline{h}_1$ with net wage $\omega_2 = W$ and virtual income $y_2 = Y$.[27]

Because the budget constraint is nonconvex, the solution cannot be characterized simply by finding a tangency with the budget constraint as it was in Section 4.6. Multiple tangencies are possible and these must be directly compared to determine the optimum. Hence, one requires the regime shift approach summarized in Section 4.4.

Consider first the regime in which positive benefits are received; that is, $h < \overline{h}_1$. Maximization, given the effective wage and income, on this linear segment follows the approach of Section 4.4. We can characterize the optimal choice according to the function $\ell(\omega_1, y_1, \nu)$. Denote the value of ν which implies $\ell(\omega_1, y_1, \nu) = 0$ as ν_0. Then the optimal hours choice along that segment is given by

$$h = \ell_1 = \ell(\omega_1, y_1, \nu),\ \nu > \nu_0;\ h = 0,\ \nu \leqslant \nu_0. \tag{4.85}$$

The optimized value on this segment (including the zero work option), accounting for the fact that $h > \overline{h}_1$ is not allowed, is given by

$$V_1^*(\omega_1, y_1, \nu) = \begin{cases} V(\omega_1, y_1, \nu), & 0 < \ell_1 \leqslant \overline{h}_1, \\ U(y_1, 0, \nu), & \ell_1 \leqslant 0, \\ -\infty, & \ell_1 > \overline{h}_1, \end{cases} \tag{4.86}$$

where Equation (4.85) defines ℓ_1.

Next, consider the regime without benefits, that is with $h \geqslant \overline{h}_1$. Again the optimal choice, given the wage and income, on this segment is given by the labor supply function $\ell_2 = \ell(\omega_2, y_2, \nu)$. The optimized value, accounting for the fact that $h < \overline{h}_1$ is not admissible, is given by[28]

$$V_2^*(\omega_2, y_2, \nu) = \begin{cases} V(\omega_2, y_2, \nu), & \ell_2 \geqslant \overline{h}_1, \\ -\infty, & \ell_2 < \overline{h}_1. \end{cases} \tag{4.87}$$

Hence, the individual selects regime 1, with welfare receipt, if $V_1^* > V_2^*$, and regime 2 otherwise. Since work propensity increases with ν, this can be characterized by a cutoff value, ν^*, defined by

$$V_1^*\big(\omega_1, y_1, \nu^*\big) = V_2^*\big(\omega_2, y_2, \nu^*\big). \tag{4.88}$$

For values of ν above ν^*, regime 2 is chosen; and for values below ν^*, regime 1 is realized.

We can define three sets, Ω_0, Ω_1, and Ω_2, such that for $\nu \in \Omega_0$ the individual chooses not to work, for $\nu \in \Omega_1$ the individual locates on segment 1 receiving benefits with positive hours of work, and for $\nu \in \Omega_2$ the individual locates on segment 2. We must consider two cases to define these sets exactly. First, suppose $\nu^* > \nu_0$. Then we have

$$\Omega_0 = \{\nu \mid \nu \leqslant \nu_0\},$$

[27] We ignore any upper bound on hours worked for simplicity.

[28] In the following formulation, we implicitly assume that the event $\ell_2 \geqslant \overline{H}$ occurs with zero probability.

$$\Omega_1 = \{\nu \mid \nu_0 < \nu \leqslant \nu^*\},$$
$$\Omega_2 = \{\nu \mid \nu > \nu^*\}. \tag{4.89}$$

Alternatively, if $\nu^* \leqslant \nu_0$, then the switch to regime 2 occurs before positive hours are worked in regime 1, that is

$$\Omega_0 = \{\nu \mid \nu \leqslant \nu^*\},$$
$$\Omega_1 = \emptyset,$$
$$\Omega_2 = \{\nu \mid \nu > \nu^*\}. \tag{4.90}$$

Hence, for certain individuals and parameter values, no value of ν exists such that they will locate on segment 1 with positive hours of work.

To characterize the likelihood function we again need a functional form for the gross wage of the form $W = W(Z) + \eta$. We ignore measurement error here for simplicity, and because there is no problem with individuals failing to locate at the kink in this nonconvex case. Define $\delta_B = 1$ if the individual receives benefits, and $\delta_E = 1$ if the individual works, both 0 otherwise. The likelihood function is given as follows, incorporating $\varphi_{\eta\nu}(\eta, \upsilon)$ and the general inverse function $\nu = \nu(h)$:

$$\begin{aligned}
&\delta_B = 1, \quad \delta_E = 1, \quad \mathcal{L}_{11} = \frac{\partial \nu}{\partial h}\varphi_{\nu\eta}\big(\nu(h), W - W(Z)\big) I(\nu \in \Omega_1),\\
&\delta_B = 0, \quad \delta_E = 1, \quad \mathcal{L}_{01} = \frac{\partial \nu}{\partial h}\varphi_{\nu\eta}\big(\nu(h), W - W(Z)\big) I(\nu \in \Omega_2),\\
&\delta_B = 1, \quad \delta_E = 0, \quad \mathcal{L}_{10} = \int_{\Omega_0} \varphi_{\nu\eta}(\nu, \eta)\, d\nu\, d\eta,
\end{aligned} \tag{4.91}$$

where $I(\cdot)$ represents an indicator function equal to 1 if the condition in the parentheses is true. Because the value of ν implied by the hours choice may be inconsistent with the value implied by the regime choice, it is possible to have "holes" in the hours distribution around the kink point. For example, an individual on segment 1 must have $\nu \leqslant \nu^*$. If his hours choice is too close to the kink, this may imply a value of $\nu > \nu^*$ and thus an observation with zero likelihood.

The overall likelihood function is given by

$$\mathcal{L} = (\mathcal{L}_{11})^{(\delta_B)(\delta_E)} (\mathcal{L}_{01})^{(1-\delta_B)(\delta_E)} (\mathcal{L}_{10})^{(\delta_B)(1-\delta_E)}. \tag{4.92}$$

Estimation proceeds by maximizing the sum of log likelihoods across individuals, as always. This is quite complex in this case, requiring knowledge of both the direct utility U and the indirect utility V, and also requiring comparisons across regimes for all individuals and all parameter values.

4.9.2. Welfare stigma implies selection of budget constraint

The above analysis assumes that all individuals eligible for welfare are on welfare. Individuals working less than h_0 but failing to receive welfare are operating below the

implied budget constraint, a possibility not permitted in the analysis. Yet, many individuals are in exactly this situation. This is generally explained by assuming the existence of some utility loss or stigma associated with welfare.

To capture welfare stigma the utility function is modified to take the form

$$U = U(c, h, \nu) - \delta_B \zeta, \tag{4.93}$$

where ζ is the level of welfare stigma which is greater than 0 and varies across individuals.[29] Two unobserved components now enter preferences, ν and ζ. Such cases were considered in the general analysis of Section 4.4. With this modification we again consider the welfare and nonwelfare regimes. Since the welfare stigma term does not affect the marginal decisions, given that the individual is on welfare, the discussion of hours of work presented above for regime 1 is still valid. The optimal utility is now given by

$$V^*(\omega_1, y_1, \nu) = \begin{cases} V_1(\omega_1, y_1, \nu) - \zeta, & 0 < \ell_1 \leqslant \overline{h}_1, \\ U(y_1, 0, \nu) - \zeta, & \ell_1 \leqslant 0, \\ -\infty, & \ell_1 > \overline{h}_1. \end{cases} \tag{4.94}$$

The analysis for regime 2 is altered in this case, because an individual can be observed not receiving welfare for any value of h – that is, given welfare stigma, it is possible to observe an individual with $h < \overline{h}_1$, but $\delta_B = 0$. So regime 2 is now defined solely by $\delta_B = 0$. Optimal hours of work, given ω_2 and y_2, are given by $\ell(\omega_2, y_2, \nu)$. Defining the value of ν for which $\ell(\omega_2, y_2, \nu) = 0$ as ν^+, hours of work under this regime are now given by

$$\begin{aligned} &h = \ell_2 = \ell(\omega_2, y_2, \nu), \quad \nu > \nu^+, \\ &h = 0, \quad \nu \leqslant \nu^+. \end{aligned} \tag{4.95}$$

Optimized utility is now

$$V_2^*(\omega_2, y_2, \nu) = \begin{cases} V(\omega_2, y_2, \nu), & \ell_2 > 0, \\ U(y_2, 0, \nu), & \ell_2 \leqslant 0. \end{cases} \tag{4.96}$$

Choice of regime still proceeds by comparing V_1^* and V_2^*, as done in relationship (4.88). For any ν in the sets Ω_0 or Ω_1 defined by expressions (4.89) or (4.90), there is now some critical level of $\zeta^* = \zeta^*(\nu)$, which depends on ν, such that regime 2 is chosen when $\zeta > \zeta^*$; regime 1 is chosen otherwise.

Given this characterization, we can derive the likelihood function for each combination of δ_B and δ_E, using the joint densities $\varphi_{\nu\zeta\eta}(\nu, \zeta, \eta)$ and $\varphi_{\nu\eta}(\nu, \eta)$:

$$\delta_B = 1, \quad \delta_E = 1, \quad \mathcal{L}_{11} = \frac{\partial \nu}{\partial h} \int_0^{\zeta^*} \varphi_{\nu\zeta\eta}\big(\nu(h), \zeta, W - W(z)\big) I(\nu \in \Omega_1)\, d\zeta,$$

[29] This additive form is used for simplicity. More general forms can be used, but change none of the substantive points presented here.

$$\delta_B = 0, \quad \delta_E = 1, \quad \mathcal{L}_{01} = \frac{\partial \nu}{\partial h}\varphi_{\nu\eta}\big(\nu(h), W - W^*(Z)\big)I(\nu \in \Omega_1) + \frac{\partial \nu}{\partial h}\int_{\zeta^*}^{\infty} \varphi_{\nu\zeta\eta}\big(\nu(h), \zeta W - W^*(Z)\big)I(\nu \in \Omega_1)\, d\zeta,$$

$$\delta_B = 1, \quad \delta_E = 0, \quad \mathcal{L}_{10} = \int_{-\infty}^{\infty}\int_{\Omega_0}\int_{0}^{\zeta^*} \varphi_{\nu\zeta\eta}(\nu, \zeta, \eta)\, d\zeta\, d\nu\, d\eta,$$

$$\delta_B = 0, \quad \delta_E = 0, \quad \mathcal{L}_{00} = \int_{-\infty}^{\infty}\int_{-\infty}^{\nu^*}\int_{0}^{\zeta^*} \varphi_{\nu\zeta\eta}(\nu, \zeta\eta)\, d\zeta\, d\nu\, d\eta. \tag{4.97}$$

Estimation proceeds as in the nonstigma case by selecting the appropriate likelihood branch for each individual and then maximizing the sum of the log likelihoods.

As with the fixed cost case, the likelihood function is complex even in this extremely simplified welfare case. For each possible set of parameter values, the maximum must be computed for each regime and then compared to compute ζ^*. Adding the tax codes, with their implied kinks, increases computational complexity. As a result, the literature has adopted a simplifying methodology which we present in Section 4.10 below.

4.9.3. Multiple program participation

In principle, the extension to the case of multiple program participation is straightforward. For simplicity, we consider a case in which the individual can choose between participating in no welfare programs, participating in welfare program 1, participating only in program 2, or participating in both welfare programs 1 and 2. We extend the utility function as follows:

$$U = U(c, h, \nu) - \delta_1\zeta - \delta_2\chi \tag{4.98}$$

where $\delta_1 = 1$ if the individual participates in program 1, and $\delta_2 = 1$ if the individual participates in program 2.[30] Benefits from program j, $b_j(I(h))$, are given:

$$b_j\big(I(h)\big) = \begin{cases} G_j - \rho_j Wh & \text{if } G_j - \rho_j Wh > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.99}$$

Benefits from both together are given as

$$b_1\big(I(h)\big) + b_2\big(I(h)\big) = \begin{cases} G_1 + G_2 - \rho_1 Wh - \rho_2 Wh = G - \rho Wh & \text{if } G - \rho Wh > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{4.100}$$

where $G = G_1 + G_2$ and $\rho = \rho_1 + \rho_2$. In general, the benefit functions for programs 1 and 2 will have different breakeven points, implying the values of hours defining kinks (H_1 in Figure 4.3) will not be the same.

[30] The use of two additive errors is a simplifying assumption which ensures that the stigma from both programs is higher than stigma from program 1 alone.

This formulation expands the model considered in Sections 4.6 and 4.9.2. To adapt this earlier model, one must designate three distinct regimes in place of regime 1 specified above: regime $1a$ indicating an individual participates only in program 1, regime $1b$ signifying this person collects benefits only from welfare program 2, and regime $1c$ designating participation in both programs. Optimal hours and utility for participation in a regime are given by (4.85), (4.86), (4.94), (4.95), and (4.96), with net wages and virtual income in these formulations specified as $\omega_j = W(1 - \rho_j)$ and $y_j = Y + G_j$, with $j = 1a$, $1b$, or $1c$. In particular, relations analogous to (4.85) and (4.86) define the labor supply and utility functions for each of the new regimes for the "on-welfare" segments associated with the relevant combination of welfare programs. Relations (4.95) and (4.96) still define the labor supply and utility functions for the nonwelfare regime. The set of relations define thresholds for ν demarcating the regions of unobserved tastes determining when a person works (ν_0 in (4.85) and ν^+ in (4.95)). Maximization again requires selection of a regime. Relations analogous to (4.94) and (4.96) characterize utilities corresponding to the various regimes. Conditional on values ν, these relations in turn imply thresholds for the stigma errors ζ, χ, and $\zeta + \chi$ that determine individuals' welfare participation. The likelihood function for this model takes a form similar to Equation (4.97), with more branches appearing in the function reflecting the additional regimes analyzed in this formulation.

Again, note the complexity of these extremely simplified welfare cases; even these involve a significant computational burden. For each possible set of parameter values, one must compute the maximum for each regime, account for the benefit structure, and then compare these to compute the error ranges for the likelihood function. When the individual is unemployed, one must perform these calculations for all possible wage values and all values of ν consistent with the no-work decision. Adding the tax code, with its implied kinks, increases computational difficulties. Introducing additional sources of unobserved heterogeneity enlarges the number of dimensions over which one must calculate integrals, requiring sophisticated numerical procedures and considerable computer resources. As a result, the literature has adopted simplifying methodologies, a topic to which we now turn.

4.10. Computational simplification by making hours choices discrete

To make estimation problems manageable, a popular method is to presume that consumers face only a limited set of hours choices. For example, a worker may choose only full-time work, part-time work, or no work, with each of these options implying a prescribed number of hours. Formally, this is done by assuming that unobservable tastes components, ν, possess a discrete distribution, usually characterized as a multinomial distribution conditional on covariates. Combined with a $0/1$ welfare decision, this finite set of hours choices yields a relatively small set of discrete states, say a set of S states, over which the utility function must be maximized.

Given a specific form for the preference function, utility can be readily evaluated at each of the hours choices and the maximum can be determined. Given an assumed joint

distribution for unobservable taste components, ν, for the error component determining wages, η, and for welfare stigma, ζ, one can compute a probability that a family selects alternative j. This in turn defines a sample log likelihood of the form

$$\mathcal{L} = \sum_{j \in S} d_j \ln P(j|X, \theta) \tag{4.101}$$

where d_j is an indicator for whether individual i chooses alternative j, X is a vector of observable characteristics, and $P(j|X, \theta)$ is the probability of choosing alternative j with θ the set of unknown parameters. Such formulations are substantially less complicated than the specifications considered above because one avoids the intricate process of calculating thresholds and dealing with combined continuous-discrete endogenous variables; only discrete choices are allowed for here.

This formulation requires each individual to be placed into a limited set of preassigned work states, even though observed hours worked take many more values, making hours look as if they were continuously distributed. To overcome this issue, analyses applying this approach necessarily introduce measurement error in hours of work to admit hours to deviate from the discrete values assumed for the choice set. Hence, conditional on ν, each alternative j contributes some positive probability $P(j|X, \theta, \nu)$ which now depends on the value of the unobservable measurement error variables.

We illustrate this approach by considering the linear measurement error model given in Section 4.3.1 where the reporting error $\varepsilon \sim \varphi_\varepsilon$, with ε and ν independent. Further, as typically assumed, we specify that hours are not subject to measurement error in no-work states. The likelihood function for hours now takes the form

$$\mathcal{L} = \left(\sum_{j \in S_0} d_j \ln P(j|X, \theta) \right)^{1-\delta_E} \left(\sum_{j \in S_1} d_j \ln\big(\varphi_\varepsilon(H - h_j) P(j|X, \theta)\big) \right)^{\delta_E} \tag{4.102}$$

where δ_E denotes a $0/1$ variable with 1 indicating that the individual works, S_0 designates the set of all states associated with the individual not working, the set S_1 includes all states in which the individual works, and h_j denotes the admissible values of true hours. Earnings depend on the values of h_j and wages. In (4.102), observed hours (H) are continuously distributed among workers.

5. Family labor supply

The study of family labor supply is motivated by a need to understand how a couple responds to tax and welfare benefit incentives when the benefit rules create links in the incentive structure as well as the need to understand how welfare is distributed within the household, so as to design the targeting of benefits appropriately. Indeed the structure of family labor supply has changed quite substantially and this may be partly due to changes in the benefit structure as well as a result of changes in relative wages. For example, in the UK there has been a large increase in the participation rate of married

women and a decrease in the participation of men. These changes have been accompanied by an increase in the number of families where no one works. This is perhaps predictable given the structure of the benefit system. However the design of income maintenance programs that target the right households and offer the right incentive system is of course important and crucially relies on our knowing the way that family labor supply is determined.

The basic family labor supply model for a married couple is the unitary model where the household is seen as maximizing one (household) utility function whose arguments are male and female labor supply and consumption. Applying demand analysis one can derive the implications of changes in wages and unearned income for behavior. Since taxes can be viewed as changes in wages and unearned income, such models can be used to simulate the labor market effects of changes in the tax system or welfare benefits. However in this model intra-household distribution has little meaning and of course the model has nothing to say about this. In addition it is unclear how the household utility function can come about from the interaction of two individuals with incentives that are not necessarily perfectly aligned. This has led to the recognition that even when dealing with households we need to account for individuals within households and we need to model the way they share resources. This leads to potentially richer models of behavior that are capable of explaining much more than the standard household model.

In the sections that follow we outline the two models and some of their implications in greater detail.

5.1. The standard 'unitary' family labor supply model

Consider the family labor supply and consumption problem

$$\begin{aligned} &\max \quad U(c, h_1, h_2, x) \\ &\text{such that} \quad c = y + w_1 h_1 + w_2 h_2 \end{aligned}$$

where U is a strictly quasiconcave function of consumption c and the two labor supplies h_i. The budget constraint equates household consumption to total income, consisting of unearned income (y) and the two earnings ($w_i h_i$), T being total time available for market work and w_i the two wages. In addition to the budget constraint, leisure cannot exceed T and hence labor supply must be positive or zero (h_i). This is a standard demand analysis problem with the complication that there may be corner solutions and wages being individual specific are not observed when the individual is working.

The first-order conditions for an interior solution simply state that the marginal rate of substitution between the two leisures will equal the ratio of wages

$$\frac{U_{h_1}}{U_{h_2}} = \frac{w_1}{w_2}. \tag{5.1}$$

An implication of this model is that behavior is neutral to within-household lump-sum redistributions of income. Thus paying a benefit to the male or the female will

have exactly the same effect, so long as it does not distort wages. This is often termed the income pooling hypothesis and we revisit the issue when we discuss the collective model. Here it suffices to note that the symmetry condition and the income pooling hypothesis are properties of the unitary model and may not be satisfied in the collective one.

5.1.1. Nonparticipation

In this subsection we show how to deal with nonparticipation and missing wages in the family labor supply context. The issues are very similar to those already discussed in the single-person labor supply model.

The first issue to be addressed is allowing for unobserved heterogeneity in the parameters of the utility function. Typically this can be addressed in a number of ways. One way would be to assume that the marginal rate of substitution for each of the two leisures with consumption includes a multiplicative error term [see Heckman (1974a, 1974b, 1974c) for example]. In this case we could write the first-order conditions as

$$\begin{aligned} \ln\left(-\frac{U_{h_1}}{U_c}\right) &= \ln(w_1) + \varepsilon_1, \\ \ln\left(-\frac{U_{h_2}}{U_c}\right) &= \ln(w_2) + \varepsilon_2. \end{aligned} \tag{5.2}$$

We can also assume a (bivariate) density for the wage rates, say $f(w_1, w_2|z)$ where z are the observable characteristics that drive wages and ε_1 and ε_2 will be taken to be independent of them. Typically one would assume a distribution function for $\varepsilon = [\varepsilon_1, \varepsilon_2]'$, for example $N(0, \Omega)$.

The functions (5.2) together with the distributional assumption for the unobserved heterogeneity define the distribution of hours of work. Hence the likelihood contribution for a couple where both are participating is simply the joint density of hours of work and wages for the two of them:

$$\begin{aligned} &\ell(h_1, h_2, w_1, w_2) \\ &\quad = |J| g\left(\ln\left(-\frac{U_{h_1}}{U_c}\right) - \ln(w_1), \ln\left(\frac{U_{h_2}}{U_c}\right) - \ln(w_2) | w_1, w_2, x\right) \\ &\qquad \times f(w_1, w_2|z), \\ &J = \frac{\partial \varepsilon}{\partial h'} \quad [\textit{Jacobian}] \end{aligned}$$

where x are observables affecting individual preferences and $h = [h_1, h_2]'$. When one or both partners are not working, hours of work are censored and the respective wage is unobserved. Take as an example the case where one of the two is not working (say 1). In this case note that $\varepsilon_1 < \ln(-\frac{U_{h_1}}{U_c}) - \ln(w_1)$, where $-\frac{U_{h_1}}{U_c}$ is the marginal rate of substitution evaluated at hours $h_1 = 0$. The likelihood contribution must be written taking this censoring into account. We will write this in terms of the joint density of

hours and wages given above. Thus the likelihood contribution for this case is

$$\ell(h_1 = 0, h_2, w_1, w_2) = \int_{w_1} \int_{h_1 \leqslant 0} \ell(h_1, h_2, w_1, w_2)\, dh_1\, dw_1.$$

The integration with respect to the wage takes place over the entire range of wages. The contributions to the likelihood for the case of the other partner not working or both not working can similarly be derived. The sample likelihood is then the product of all contributions. In a similar fashion one can construct the likelihood contribution for the case where neither member of the household is working. The sample likelihood is then the product of the contributions for each observation. This is the basic likelihood structure. We next discuss issues relating to introducing taxation in this framework.

5.2. *Discrete hours of work and program participation*

It is straightforward to allow for proportionate taxes, or even piecewise-linear taxes, so long as these lead to a budget constraint that is convex and so long as the endogeneity of the tax rate is taken into account. However, most welfare programs are designed in such a way that they define a nonconvex budget set: Implicit marginal tax rates are higher at low hours of work, where increases in earnings lead to a rapid withdrawal of benefits, and lower at higher hours where the individual pays the usual taxes. As we showed earlier, this is a complex problem itself and in the family labor supply context even more so because the benefits may be interdependent.

To simplify the problem it has now become almost standard to discretize hours of work. Then the problem of utility maximization becomes one of choosing packages of consumption and earnings – consumption is defined by the earnings of the individual, the tax system and the benefit system. Within this context we can also account for fixed costs of work (another nonconvexity) and for the decision to participate (or not) in a welfare program [Hoynes (1996) and Keane and Moffitt (1998)].

We start with a utility function defined over hours of work H_1 and H_2 and we discretize the distribution of hours. For example hours can take the discrete values $H = \{0, 20, 40\}$. Suppose we write family utility at hours $H_1 = h_i$, $H_2 = h_j$ where h_i and h_j are the ith and jth point of the discrete hours distribution respectively:

$$U_{h_i h_j} = U(H_1 = h_i, H_2 = h_j, c, \varepsilon) - \eta P_B + u_{h_i h_j}$$

where P_B is a 0–1 program participation dummy. The term ηP_B reflects the utility costs of program participation such as "stigma". This may be randomly distributed over the population. The term ε reflects unobserved heterogeneity in preferences and the term $u_{h_i h_j}$ hours-specific unobserved heterogeneity. Given the associated wage, the discrete hours imply a corresponding set of earnings for each individual.

The budget constraint incorporates all relevant aspects of the tax and benefit system to define the resulting level of consumption

$$c = w_1 H_1 + w_2 H_2 + y - T(y, w_1 H_1, w_2 H_2) + B(y, w_1 H_1, w_2 H_2) P_B$$

where T is the tax function and B is the program benefit function.

The likelihood is derived taking into account program participation. First note that the observation on whether an individual is participating in welfare programs or not is informative about the range in which the participation cost η lies. Note also that for any given η the utility function and the budget constraint define whether the person will be a participant or not. At each observation we can derive the probability $\text{pr}(U_{h_i h_j} > U_{h_k h_s}, \forall k \neq i \text{ and } s \neq j | w_1, w_2, \eta, \varepsilon)$ that the chosen point is optimal, conditional on η and the heterogeneity terms ε. If the person is eligible for a welfare program at the observed point and he does actually participate (i.e. receives benefits) then the range in which η lies is defined by the fact that the utility gain from participating is higher than the cost η. For the nonparticipants η lies in the complement of this set. This allows us to integrate out η over the relevant range. When the person is ineligible at the observed point no information is available for η and we integrate over its entire range. In this case as we move over different values of η the probabilities change not only because of the direct effect of η through the utility function but also because it induces different potential participation decisions at each discrete hours point, thus changing both optimal hours and consumption. Thus consider the likelihood contribution for a couple where both work and participate in a welfare program (in-work benefits). This will take the form

$$\left\{ \int_{\eta \epsilon Q} \int_{\varepsilon} \text{pr}(U_{h_i h_j} > U_{h_k h_s}, \ \forall k \neq i \text{ and } s \neq j | w_1, w_2, \eta, \varepsilon) \, d\varepsilon \, d\eta \right\} f(w_1, w_2 | z)$$

where Q is the set of η such that program participation is optimal at the point of observation. The form of the probability is defined by $U_{h_i h_j}$. Imposing a logistic is not restrictive if we allow for unobserved heterogeneity through the ε [Manski and McFadden (1981)]. The contribution to the likelihood for a nonworker must also take into account the fact that the wage will not be observed in that case. This is done as before by integrating over all possible wages. Of course the practical difficulty is that the probability of participation is a complicated function of the wage rate through the formulae of the tax and welfare benefit system.

The models estimated in this way have the great attraction that they allow us to simulate policies allowing for possible changes in the take-up of means-tested benefits. To the extent that there is sufficient genuine exogenous variation in the data to allow us to identify the factors that determine take-up these can be very useful for the *ex ante* evaluation of welfare policies.

5.3. *Collective models of family labor supply*

The family labor supply model presented above treats the household as a single optimizing decision unit, and has nothing to say about within-household allocations. It also imposes stronger restrictions than necessary, such as symmetry. An alternative approach, the collective model, looks upon the household as a set of individuals with their own preferences, who have to decide how to share the overall set of resources available to them. Within this framework we can have private goods (enjoyed by the members

separately), public goods and household production.[31] The main empirical issue is that of identification: What can we learn about individual preferences and the sharing rule when we observe aggregate household consumption. This has led to a number of important theoretical results by Chiappori (1988, 1992) recently extended by Blundell et al. (2007) to allow for corner solutions and to discuss identification in the presence of unobserved heterogeneity.

The framework we describe here is the collective model with two household members and no public goods or household production.[32] Each member supplies labor h^i $(i = m, f)$ and consumes a private good (C^i). A critical assumption in the collective approach as introduced by Chiappori is that the household only takes Pareto-efficient decisions. That is, for any set of male and female wages and unearned income (w_f, w_m, y), there exists some level of male utility $\overline{u}^m(w_f, w_m, y)$ such that labor supply and consumption for each household member (h^i, C^i) is a solution to the program:

$$\begin{aligned}
&\max_{h^f, h^m, C^f, C^m} \quad U^f\left[1 - h^f, C^f\right],\\
&U^m\left[1 - h^m, C^m\right] \geqslant \overline{u}^m(w_f, w_m, y),\\
&C = w_f.h^f + w_m.h^m + y,\\
&0 \leqslant h^i \leqslant 1, \quad i = m, f,
\end{aligned} \tag{5.3}$$

where the labor supply has been normalized to lie between 0 and 1. The function $\overline{u}^m(w_f, w_m, y)$ defines the level of utility that member m can command when the relevant exogenous variables take the values w_f, w_m, y. Underlying the determination of $\overline{u}^m$ is some allocation mechanism (such as a bargaining model) that leads to Pareto-efficient allocations. The nice thing about the collective approach is that there is no need to be explicit about such a mechanism; identification does not rely on specific assumptions about the precise way that couples share resources.

Suppose first that preferences are such that there are never any corner solutions. It is assumed that we observe aggregate household consumption $C = C^m + C^f$ and that we know the locus of labor supplies as a function of (w_f, w_m, y). Then Chiappori (1988) proves the following:

PROPOSITION 5.1. *[See Chiappori (1988).] Assume that h^m and h^f are twice differentiable functions of wages and nonlabor income. Generically, the observation of h^m and h^f allows us to recover individual preferences and individual consumptions of the private good up to an additive constant.*

[31] There have also been a number of tests of the unitary model, typically rejecting it and motivating work on collective models. These papers include Thomas (1990), Fortin and Lacroix (1997), Browning and Chiappori (1998).

[32] Blundell, Chiappori and Meghir (2005) further extend the model to discuss identification conditions with public goods.

There are two critical issues to be resolved following this proposition: One is what happens with corner solutions and with discrete labor supply. The other is what happens with unobserved heterogeneity in preferences, i.e. when we do not know the exact loci h^m and h^f.

Blundell et al. (2007) set up a framework where the male decision is discrete (work or not) and the female is continuous – however she can choose not to work. The framework underlying the proposition above exploits the fact that the marginal rates of substitution between consumption and labor supply for each agent will be equalized within the household, under efficiency. This result cannot help when one of the labor supplies is discrete. Define the participation frontier to be the set of male and female wages and unearned income y so that member m is indifferent between working and not working. Blundell et al. (2007) then exploit the following implication of efficiency:

DEFINITION AND LEMMA DI *(double indifference). The participation frontier L is such that member m is indifferent between participating or not. Pareto efficiency then implies that f is indifferent as well about whether m participates or not.*

Technically, this amounts to assuming that in the program above, $\overline{u}^m$ is a *continuous function* of both wages and nonlabor income. This will imply that the behavior of the female will depend on the male market wage even when he is not working. This continuity assumption restricts the set of possible behavior and plays a key role for identification. We will not go through the technical details, all of which are available in the paper referenced above. However, identification of preferences and the consumption sharing rule (up to an additive constant) follows from the assumption that all goods are private (no public goods and no household production) as well as from the assumption above. Blundell, Chiappori and Meghir (2005) discuss results in the presence of public goods. The essence of the results there is that full identification of preferences over private and public goods and the sharing rule follows when preferences over private consumption and labor supply are weakly separable from the public good. In any case it is shown that some aspects of the public good must be observable.

The next important obstacle for identification here is unobserved heterogeneity. The results outlined above relate to the case where we know the locus of the observable endogenous variables (labor supplies, the public good, etc.) as functions of wages and unearned income. However for empirical purposes we need to establish identification in the presence of unobserved heterogeneity in preferences. This is generally complicated by the fact that any unobserved components affecting individual preferences are likely to affect the sharing rule. Since this can take any form (more or less) we may well end up with error terms that are nonseparable, which of course may lead to lack of identification in general. Identification problems are compounded by the specific context of labor supply where wages are only observed for workers. Blundell et al. (2007) have established identification in the special case where the labor supplies and the sharing rule are linear in log wages and all have additive unobservables. Even in this case the proof is not trivial because they do not rely on distributional assumptions. One conclusion of

this study is that identification in more complex preference structures will have to be established on a case-by-case basis. Nevertheless, the dividends of such an exercise are probably very high. Blundell et al. (2007) reject the unitary model, while the collective model is not rejected and gives interesting insights into the way that resources are split up within the household. Further empirical work needs to include public goods and household production. This will allow an extension of this analysis to households with children. Finally, this framework needs to be extended to deal explicitly with the issues of taxation and means-tested benefits, which the previous analysis of the collective model has not developed.

6. Intertemporal models of labor supply

The models discussed up to now focused on the work decision within a period. The life-cycle and dynamic issues have not been addressed. However, studying dynamics is of critical importance because of the numerous intertemporal dependencies in labor supply and their implications for the design of policy.

The most obvious intertemporal dependence comes through borrowing and saving. In this framework the credit market is used to shift labor income across periods of the life-cycle so that labor supply can be concentrated in periods when the relative benefit of supplying labor is highest or costs are lowest. This allows a reduction in labor supply during college, during childrearing and during retirement while consumption can be maintained at a level consistent with expectations and overall uncertainty. An additional reason for changes in labor supply over the life-cycle is the precautionary motive, which implies more labor supply when one is young and less when one is older and some of the uncertainty has been resolved [Low (1999)].

However, intertemporal dependence may be more direct. Labor supply preferences may depend on past actions (habit formation); current work may improve future wages through learning by doing; current work may increase a future pension entitlement. Since a rational individual will take into account the impact of current actions on future budgets or preferences, the standard static labor supply model does not tell the complete story and may in fact be misleading. With intertemporal dependencies the individual may find it rational to work in circumstances where the static model would exclude such a possibility. For example, it may still be worth working when welfare benefits are reduced one for one with earnings, because work offers future returns in the form of higher wages.

The recent intertemporal labor supply literature has developed along two lines. This is reflected in these two intertemporal aspects of labor supply – through credit markets and saving, and through intertemporal nonseparabilities. In the former case applications exploit the continuity of consumption and saving to derive Euler equation conditions for intertemporal labor supply. In the latter case the focus is more on participation and intertemporal nonseparabilities, largely ignoring saving decisions.

This classification of approaches is necessarily too restrictive. There are intertemporal substitution applications that allow nonseparability over time, but these are few and

typically do not account for fixed costs and nonparticipation. Also there are examples of dynamic programing models that account for saving decisions but to date these have been quite rare and based on very specific assumptions concerning preferences and markets.

This section presents dynamic models of labor supply and consumption and discusses their estimation. We start by presenting the standard dynamic framework, followed by the empirical models of MaCurdy (1981) and Heckman and MaCurdy (1980). We then discuss issues to do with intertemporal nonseparability, and unobserved heterogeneity in the context of incomplete insurance markets. We conclude with the presentation of a framework in which all these aspects are taken into account in a theoretically coherent fashion.

6.1. Intertemporal labor supply with saving

As we have mentioned in Section 2, the "static" labor supply model can be made consistent with an additively separable life-cycle model under uncertainty using the two-stage budgeting framework. However, this does not recover all of the parameters necessary for intertemporal analysis and for that we need to look directly at the first-order conditions for intertemporal optimization. Before moving to consider the problems of unobserved heterogeneity in the context of uncertainty and with the possibility of corner solutions we consider a simpler model.

Using the framework of Heckman and MaCurdy (1980) and MaCurdy (1981) we discuss estimation of life-cycle labor supply models in a complete markets setting, i.e. with no uninsurable uncertainty and no aggregate shocks. We start by exposing the case of no corner solutions, where all individuals work. We then allow for nonparticipation. Next we introduce uncertainty, first by considering the no corners case and later allowing for corners as well. Finally we discuss the issue of unobserved heterogeneity in models with uncertainty and corner solutions and present an estimation framework based on the complete dynamic programing characterization of the problem.

6.1.1. The life-cycle model

Before discussing the identification and estimation issues in the dynamic models of labor supply and consumption we present the standard life-cycle model.[33]

The individual maximizes expected lifetime utility subject to an intertemporal budget constraint. We assume that future wage rates, prices and interest rates are uncertain and that labor market risk is uninsurable. Define A_t to be the assets, denominated in the same units as consumption. Letting i_t denote the nominal interest rate and p_t the price level, we define the real rate of return on assets to be $1 + r_t = \frac{p_t}{p_{t+1}}(1 + i_t)$. Thus r_t is to be taken as uncertain in period t. The real wage rate is denoted by w_t.

[33] We draw from Browning, Deaton and Irish (1985) and Blundell, Browning and Meghir (1994).

Denote by E_t the expectations operator with respect to the distribution of uncertain future variables conditional on information in period t. These include interest rates, wages, the price level, possible preference shocks and other variables which affect choices either through their impact on expectations or directly. Denote the collection of such *state* variables by S_t. The state variables contain all the information that is needed to summarize the individual's position at any point in time. Thus, conditional on the state variables the past otherwise is irrelevant. We can also think of the taste shifter variables z_{1t} and z_{2t} as being uncertain in future periods, in which case expectations are taken with respect to their distribution as well. We abstract from issues relating to uncertain date of death and the presence or absence of perfect annuity markets. Hence we take the personal discount factor β to be constant over time as a simplifying assumption.

We can write the intertemporal optimization problem as

$$V_0 = \max_{h_t, c_t} \left\{ E_0 \sum_{t=0}^{T} \beta^t \psi\big[U(c_t, h_t | z_{1t}), z_{2t}\big] \,\Big|\, \sum_{t=0}^{T} \frac{1}{\prod_{s=0}^{t}(1+r_s)} (c_t - w_t h_t) \geqslant 0 \right\}$$

where the second part in the expression is the intertemporal budget constraint. The way it is written implies that the individual can borrow and lend freely at a market rate of interest r_t.

The additive structure of this problem is viewed from the perspective of period 0. However, since there exists uninsurable uncertainty the individual will replan in each period as news arrives. In this context and since the problem is recursive (trivially since it is additive over time) it is more convenient to use the Bellman equation formulation

$$V_t(A_t | S_t) = \max_{h_t, c_t} \big\{ \psi\big[U(c_t, h_t | z_{1t}), z_{2t}\big] + E_t \beta V_{t+1}(A_{t+1} | S_{t+1}) \big\} \tag{6.1}$$

where $V_t(A_t | S_t)$ is the optimum value function given information up to period t and S_t are relevant state variables which help predict future uncertain income, interest rates and characteristics.

In the absence of credit market restrictions the intertemporal budget constraint implies that

$$A_{t+1} = (1 + r_t)(A_t + w_t h_t - c_t)$$

with the terminal value of assets fixed at some value (say zero).[34] This implies that the revenues and expenditures need to balance over the entire life-cycle but not necessarily at any point in time.

The first-order conditions for labor supply and consumption can be written as

$$-u'_h \geqslant \lambda_t w_t, \quad h_t \geqslant 0,$$
$$u'_c \geqslant \lambda_t, \quad c_t \geqslant 0.$$

[34] We abstract from issues relating to portfolio choices and r_t is the return to the market portfolio.

Usually an Inada condition is imposed which ensures that optimal consumption will always be strictly positive. However, optimal labor supply may be zero which leads to a corner solution.

For individuals with an interior solution the optimal allocation between consumption and hours of work within a period equates the marginal rate of substitution to the real wage rate. The important point is that even in this dynamic context the marginal rate of substitution is the ratio of *within-period* marginal utilities. Thus consumption and labor supply satisfy

$$-\frac{u'_h}{u'_c} = w_t \tag{6.2}$$

where u'_x is the marginal utility of x. The important point to note is that the within-period marginal rate of substitution between consumption and hours of work does not depend directly on any expectations about the future, nor does it depend on interest rates.[35] Crucially, it does not depend on the monotonic transformation of the utility function ψ. This is important because it implies that in general we cannot estimate the parameters governing intertemporal allocations just by using within-period ones. Condition (6.2) is the basis of the life-cycle consistent "static" labor supply model of the earlier sections.

We can apply the envelope condition for assets on (6.1) to characterize the link between decisions over time. This gives

$$V'_t = E_t\{\beta(1+r_t)V'_{t+1}\}.$$

Since the first-order conditions also imply that

$$\psi'_t U'_{ct} = E_t\{\beta(1+r_t)V'_{t+1}\}$$

and

$$\psi'_t U'_{ht} = -E_t\{\beta(1+r_t)w_t V'_{t+1}\}$$

we can characterize the intertemporal rates of substitution for consumption and hours of work for interior solutions as

$$\psi'_t U'_{ct} = E_t\{\beta(1+r_t)\psi'_{t+1}U'_{ct+1}\}, \tag{6.3}$$

$$\psi'_t U'_{ht} = E_t\left\{\beta(1+r_t)\frac{w_t}{w_{t+1}}\psi'_{t+1}U'_{ht+1}\right\}. \tag{6.4}$$

The object of the exercise is to estimate the parameters of $\psi[U(c_t, h_t|z_1), z_2]$ from observations of consumption and labor supply over time. It turns out that we need to use two of the three conditions (6.2), (6.3) and (6.4). At this point note that the variables in

[35] This important point has been made by among others MaCurdy (1983), Blundell and Walker (1986), Altonji (1986) and Arellano and Meghir (1992).

z_1 affect both the within-period marginal rate of substitution and intertemporal allocations. The z_2 variables only affect directly intertemporal allocations because they cancel out of the monotonic transformation. Of course they do affect within-period allocations indirectly and in a full solution the consumption and labor supply functions will depend on all variables affecting tastes, expectations and the budget.

6.1.2. A simplification: A model with full participation

Before complicating matters with nonparticipation we consider the estimation problem in a simpler model presented by MaCurdy (1981) where everybody works. The utility specification he used does not allow for corner solutions and takes the form

$$U_t = B_t c_t^{\gamma} - A_t H_t^{\alpha}, \quad 0 < \gamma < 1,\ \alpha > 1, \tag{6.5}$$

where H_t corresponds to hours of work (rather than leisure) and C_t to consumption. The range of parameters ensures positive marginal utility of consumption, negative marginal utility of hours of work and concavity in both arguments. Applying exactly the same analysis as above the implied intertemporal Frisch labor supply becomes

$$\ln H_t = A_t^* + \frac{1}{\alpha - 1} \ln \lambda + \frac{1}{\alpha - 1} \ln w_t + \frac{\rho - r}{\alpha - 1} t \tag{6.6}$$

where the use of log hours of work presumes that all individuals work and hence $H > 0$. In (6.6) λ is the shadow value of the lifetime budget constraint and t is the age of the individual. Finally A_t^* reflects preferences and is defined by $A_t^* = -\frac{1}{\alpha - 1} \log A_t$.

This equation is the Frisch labor supply equation. The important insight is that under certainty (complete markets – no aggregate shocks) all relevant future variables, such as wages are summarized by the fixed effect λ. So this equation has a simple message: Hours of work are higher at the points of the life-cycle when wages are high ($\frac{1}{\alpha - 1} > 0$). Moreover if the personal discount rate is lower than the interest rate, hours of work decline over the life-cycle. Finally, hours of work will vary over the life-cycle with A_t^*, which could be a function of demographic composition or other taste shifter variables.

Specifying $A_t^* = \gamma' x_t + \eta_1 + u_t$ we obtain an econometric equation of the form

$$\ln H_t = \gamma' x_t + \frac{1}{\alpha - 1} \ln w_t + \frac{\rho - r}{\alpha - 1} t + \left[\frac{1}{\alpha - 1} \ln \lambda + \eta_1 \right] + u_t \tag{6.7}$$

where $[\frac{1}{\alpha - 1} \ln \lambda + \eta_1]$ is a fixed unobservable individual effect consisting of the marginal utility of wealth and of a permanent unobserved preference component. u_t is an idiosyncratic shock to individual preference. For simplicity we take this as serially uncorrelated.

As it is, this equation presents a problem for estimation to the extent that the fixed unobservable effect (or the idiosyncratic shock u_t) is correlated with the hourly wage rate w_t. Because λ is a function of all wages over the life-cycle and because wages are highly persistent it is not tenable to assume that the fixed unobservable is not correlated

with wages. The simplest case here is to assume that all right-hand-side variables, including wages, are strictly exogenous, namely that $E(u_t|x_s,\ \ln w_s, \forall s = 1, \ldots, T) = 0$ in which case the model can be estimated using within-groups estimation: variables are transformed into deviations from their individual-specific time mean and OLS is applied on

$$\widetilde{\ln H}_t = \gamma' \tilde{x}_t + \frac{1}{\alpha - 1}\widetilde{\ln w}_t + \frac{\rho - r}{\alpha - 1}\tilde{t} + \tilde{u}_t \tag{6.8}$$

where $\tilde{z}_t = z_t - \bar{z}$ represents the deviation of an individual-specific variable from the time mean for this individual. This model is estimable using panel data with a relatively small number of repeated observations for each of many individuals.[36] Here Ordinary Least Squares on the transformed model is consistent and fully efficient.

This empirical strategy is sensitive to measurement error for the right-hand-side variables. Suppose that log wages are measured with additive and serially uncorrelated (classical) measurement error. In this case the strict exogeneity assumption is violated and (6.7) cannot be estimated by within groups. An alternative approach in this case would be to take first differences, thus eliminating the fixed effect and then using instrumental variables to estimate the parameters based on the transformed equation. The instruments would have to be dated t–2 or earlier because the error in the first difference equation will have an MA(1) structure. Thus, under the assumptions made, valid instruments would be hours and wages lagged at least two periods. However, these instruments will only be valid if they are able to explain future growth in wages ($\Delta \log w_t$); hence this rank condition needs to be tested.

6.1.3. The Heckman and MaCurdy study

The MaCurdy (1981) paper set out the first clear analysis of issues to do with estimating intertemporal labor supply relationships. However the approach did not deal with corner solutions, which is particularly relevant for women. The first attempt to do so in the context of a life-cycle model of labor supply and consumption is the paper by Heckman and MaCurdy (1980). In this model women are endowed with an explicitly additive utility function for leisure L and consumption c in period t, of the form[37]:

$$U_t = A_t \frac{L_t^{\alpha} - 1}{\alpha} + B_t \frac{c_t^{\gamma} - 1}{\gamma}, \quad \alpha, \gamma < 1. \tag{6.9}$$

Consumers are assumed to maximize life-cycle utility

$$V_t = \sum_{t=1}^{T} \beta^t U_t$$

[36] Fixed T and large N asymptotics.

[37] See also Altonji (1982).

subject to the lifetime budget constraint

$$\sum_{t=1}^{T} \frac{1}{(1+r)^t}[w_t h_t - c_t] \geqslant 0$$

where $h_t = \overline{L} - L_t$, $\overline{L}$ being maximal time available for work, and where w_t is the hourly wage rate. Note that now utility depends on leisure and is well defined at the point where hours are zero since there one obtains maximum leisure.

Optimization is assumed to take place under perfect foresight. Solving for the first-order conditions we obtain the following equation for leisure:

$$\ln L_t = \begin{cases} A_t^* + \frac{1}{\alpha-1}\ln w_t + \frac{\rho-r}{\alpha-1}t + \lambda^* & \text{when the woman works,} \\ \ln \overline{L} & \text{otherwise,} \end{cases} \tag{6.10}$$

where

$$\lambda^* = \frac{1}{\alpha-1}\ln\lambda \quad \text{and} \quad A_t^* = -\frac{1}{\alpha-1}\ln A_t \tag{6.11}$$

and where we have approximated $\ln\frac{1+\rho}{1+r} \approx \rho - r$. Note that in contrast to specification (6.6), the parameter α is defined to be less than unity. As before in (6.11) λ is the shadow value of the lifetime budget constraint which again is a fixed effect because of the complete markets assumption and t is the age of the individual.

6.1.3.1. Estimation with nonparticipation To estimate the model, Heckman and MaCurdy specify $A_t^* = \gamma' x_t + \eta_1 + u_{1t}$ where u_{1t} is normally distributed and where η_1 is a fixed effect reflecting permanent unobserved differences in tastes across individuals.

Given λ^*, η_1 and wages w_t this gives rise to a Tobit model, with censoring whenever the interior solution requires more hours of leisure than are available ($L_t > \overline{L}$). There are two main difficulties with this however. First, hourly wage rates are not observed for nonworkers. Second, λ^* and η_1 are unobserved and cannot be differenced out in a conventional manner since the Tobit model is essentially nonlinear. Finally, a problem addressed only indirectly before (through the treatment of measurement error) is that of the endogeneity of wages. To solve these problems and to take into account that wages may be endogenous we may specify a wage equation of the form

$$\ln w_t = z_t'\beta_2 + \eta_2 + u_{2t}$$

with η_2 being an unobserved fixed effect reflecting permanent productivity characteristics of the individual and u_{2t} being normally distributed. Endogeneity may arise if either the fixed effects in the wage and labor supply equations are correlated or if the idiosyncratic components are correlated (or both). In the former case (correlated fixed effects) treating the problem of fixed effects will also solve the endogeneity problem. In this sense we can think of wages as being endogenous in the case where we dealt with no corner solutions.

To proceed we can use the approach described earlier in the context of the static labor supply models. The wage equation is substituted into the structural labor supply

equation and the conditions for an interior solution or otherwise are given in terms of the reduced form, i.e. not conditional on the wage rate. Hence we get

$$\ln L_t = \begin{cases} \gamma' x_t + \frac{1}{\alpha-1} z_t' \beta_2 + \frac{\rho-r}{a-1} t + f + \upsilon_t \\ \quad \text{when } \upsilon_t < \ln \overline{L} - \left(\gamma' x_t + \frac{1}{\alpha-1} z_t' \beta_2 + \frac{\rho-r}{a-1} t + f\right), \\ \ln \overline{L} \quad \text{otherwise,} \end{cases}$$

where $f = \lambda^* + \eta_1 + \frac{1}{\alpha-1}\eta_2$, $\upsilon = u_{1t} + \frac{1}{\alpha-1} u_{2t}$. This gives rise to a Tobit model for the reduced form parameters. However, two important difficulties need to be addressed. The first relates to estimating this reduced form and the second to recovering the structural parameters characterizing labor supply.

The reduced form labor supply includes a fixed effect f. In a linear model and with strict exogeneity the within-groups estimator is consistent and efficient. The model here is nonlinear because of censoring. Heckman and MaCurdy (1980) treated the fixed effects as parameters to be estimated. Formally speaking, when the model is nonlinear, this estimator is not consistent as the number of individuals N grows, while the number of time periods per individual T remains fixed. This is because the number of (incidental) parameters grows with the sample size. In practice the estimator is likely to work well with strictly exogenous regressors for moderate to large T. Heckman and MaCurdy provide Monte Carlo evidence showing that in their context the bias involved when using this approach is likely to be minimal for moderate T. However, this is not a general result and it depends very much on the model, the data and the number of time periods available. For example with lagged endogenous variables the biases could be substantial. Such lagged endogenous variables could appear in time nonseparable models and in models with incomplete insurance markets as we will see subsequently. Thus the complete markets assumption turns out to be particularly powerful as far as identification is concerned.

An alternative approach is to use a semiparametric LAD estimator introduced by Honore (1992). This estimator relies on symmetry of the difference of the errors $(u_{it} - u_{it-1})$ conditional on the sum of the errors $(u_{it} + u_{it-1})$ and on the regressors, which is weaker than the assumption of normality combined with iid errors.

We have described how the reduced form labor supply equation can be estimated. This does not provide the parameters of the structural model because they are a function of the parameters of the wage equation. The next step is to recover the structural parameters. The difficulty here is that we first need to identify the parameters of the wage equation. This is not a simple problem because wages are observed for workers only, who are endogenously selected. In addition both the selection mechanism and probably the wage equation depend on fixed effects. Before we discuss estimation first we need to ensure that the parameters are identified. A necessary condition is that the wage equation includes variables that are excluded from the structural labor supply equation. Under normality no further restrictions are required. However, if one applies a semiparametric estimation framework that relaxes the normality assumption one also requires variables included in the labor supply equation that are excluded from the wage

equation. One approach to completing estimation is to apply the Kyriazidou (1997) estimator to the wage equation. This controls for selection allowing for fixed effects in both the wage and the participation equations. Once the parameters of the wage equation have been recovered, one can use minimum distance to back out the parameters of the labor supply equation, which were estimated as above.

An alternative approach, and one followed by Heckman and MaCurdy, is to use maximum likelihood treating the fixed effects as parameters to be estimated jointly (as discussed above). We turn to this approach now.

6.1.3.2. Maximum likelihood estimation The first step is to specify the joint distribution of hours of work and wages, conditional on the observables and the fixed unobserved effects. This is denoted by

$$g_{hw}(h, w|z, f, \eta) = g_h(h|x, f, w)g_w(w|z, \eta_1) \tag{6.12}$$

where z are the observed variables in the wage equation, which include all those in the labor supply equation (x) and more for identification purposes. In the above equation $g_h(h|x, f, w)$ is the conditional density of hours of work given wages, x, and f and $g_w(w|z, \eta_1)$ is the conditional distribution of wages given z and η_1. Thus the model likelihood is bivariate including that of wages.

The likelihood has the general form

$$\mathcal{L} = \prod_{workers} g_h(h|x, f, w)g_w(w|z, \eta_1) \cdot \prod_{nonworkers} \int_{h<0} \int_w g_h(h|x, f, w)g_w(w|z, \eta_1)\, dw\, dh. \tag{6.13}$$

The first part of the likelihood relates to workers, where both wages and hours are jointly observed. The second part of the likelihood refers to nonworkers where all we know is that desired hours are negative. Hence we integrate over $h < 0$ and over the entire support of the wage distribution, since for any wage rate there is a configuration of unobservables that would make the person a nonparticipant – being a nonworker conveys no information about wages. This likelihood can recover the parameters in the reduced form labor supply equation and in the wage equation.

As mentioned above, to identify the structural parameters of labor supply and the wage equation it is necessary to impose exclusion restrictions or some other form of parametric restrictions. Moreover, note that any variables that are fixed cannot be used for identification since they will be absorbed by the fixed effect. Heckman and MaCurdy exclude education/age interactions and aggregate unemployment from the labor supply equation and husband's labor market behavior from the wage equation. The former restriction effectively implies that differences in tastes across education groups *vis à vis* labor supply do not change with age. Consequently any change in observed behavior across education groups at different ages is attributed to education-specific changes in individual productivity and hence to wages. The business-cycle indicator (the unemployment rate) serves to identify wages for the nonworkers through the aggregate price

of human capital. Note, however, that given the functional form assumptions the model is then overidentified.

The Heckman and MaCurdy model presented above offers a way of handling unobserved heterogeneity and corner solutions and even allows for persistent heterogeneity and endogenous wages. These properties have been delivered at a cost. Preferences between consumption and female hours are explicitly additively separable and no uncertainty is allowed for. The explicit additivity implies that, given consumption data, all parameters could be identified in principle using just within-period allocations. This is worrying since it implies that intertemporal allocations are tied to the way that resources are allocated within period – an implication that does not come from economic theory. However, this assumption is testable since we can compare the estimates obtained from data on within-period and data on intertemporal allocations. Finally, the perfect foresight assumption which is equivalent to complete markets with no aggregate shocks is also strong given the available evidence.

However, easy as it may be to criticize such an approach, it turns out that it is very hard to generalize. In what follows we discuss how the existing literature has attempted to build on this and what are the successes and shortcomings of these attempts. We start by describing an estimation strategy for a model of consumption and labor supply with corner solutions but with no explicit treatment of unobserved heterogeneity. As we argue below, an explicit treatment of unobserved heterogeneity places extensive requirements on data and an approach based on the complete solution of the life-cycle model, rather than on Euler equations.

6.1.4. *Estimating the intertemporal substitution elasticity and other preference parameters under uncertainty*

We now consider explicitly estimation in the presence of uninsurable uncertainty.[38] Estimation will be based on two marginal conditions: One defines the within-period allocations and the other the intertemporal allocation. Combining these two conditions in a suitable way can allow us to identify all parameters while accounting for corner solutions.

We start by characterizing within-period preferences using the indirect utility function and appealing to two-stage budgeting. The within-period indirect utility function is defined by

$$\psi\big[v_t(w, y)|z_t\big] = \max_{h,c}\big\{\psi\big[U_t(h, c)|z_t\big] \,\big|\, c_t = w_t h_t + y_t\big\} \tag{6.14}$$

where the variables z_t are shown explicitly to emphasize that intertemporal allocations will typically depend on taste shifter variables. As explained earlier in the chapter, the variable y_t reflects net saving or dissaving. Because c_t is realized consumption and $w_t h_t$ are actual earnings this amount (y_t) will only equal unearned income (e.g. from

[38] See Blundell, Browning and Meghir (1994).

transfers or income from investments) if there is neither borrowing nor saving by the individual. Based on Roy's identity it is possible to derive the implied within-period (or Marshallian) labor supply function, i.e.

$$h(w, y) = \frac{\partial v/\partial w}{\partial v/\partial y}. \tag{6.15}$$

This labor supply function is conditional on y_t which reflects intertemporal decisions.

The labor supply function originating from (6.15) can be estimated using the methods described in earlier sections. The estimation of the within-period labor supply function allows us to estimate all the parameters characterizing within-period preferences, i.e. the function $v_t(w, y)$ in (6.14) but not the parameters of the function ψ. The latter affects intertemporal allocations only.

Thus we now need data on intertemporal allocations to estimate the parameters implicit in the monotonic transformation ψ, which characterizes saving behavior and intertemporal substitution in labor supply.

Consider again the Euler equation in an environment with uninsurable risk. This equates the marginal utility of consumption today with the expected marginal utility of consumption tomorrow:

$$\psi_t' v_{yt}' = E_t\{\beta(1 + r_t)\psi_{t+1}' v_{yt+1}'\}.$$

The term $v_{yt}' = \frac{\partial v}{\partial y}$ is the marginal utility of money, and $\psi_t' = \frac{\partial \psi[v_t(w,y)|z_t]}{\partial v_t}$ reflects the monotonic transformation of the utility function, which determines the intertemporal substitution. The marginal utility of money v_{yt}' can be estimated as a first step based on observations relating to within-period allocations. We denote the estimated quantity by $\hat{v}_{yt}'$. The next step is to parameterize the function ψ which can then be estimated using the Euler condition. To write the Euler condition based on the indirect utility function we can use the envelope theorem to see that $U_{ct}' = v_{yt}'$ where U_{ct}' is the marginal utility of consumption which appears in the Euler condition (6.3). Based on this we can estimate the parameters characterizing ψ_t' using the following equation:

$$\psi_t' \hat{v}_{yt}' = \beta(1 + r_t)\psi_{t+1}' \hat{v}_{yt+1}' + u_{it+1} + \varepsilon_{it} \tag{6.16}$$

where ε_{it} represents the estimation error due to the fact we are replacing v_{yt}' with its estimated value. Under the hypothesis of rational expectations any variable dated t or earlier will be orthogonal to u_{it+1}. This observation can serve as a basis for estimation using GMM [see Hansen (1982) and Hansen and Singleton (1982)]. Asymptotically ε_{it} will become irrelevant if the first-step estimator is consistent, but it can have serious implications in small samples.

With uninsurable uncertainty and in the presence of aggregate shocks it is imperative to estimate (6.16) using long enough time series. The innovation to the marginal utility of wealth u_{it+1} reflects uninsurable idiosyncratic risk and aggregate uncertainty. As Altug and Miller (1990, 1998) have shown, the moment conditions do not hold in the cross section. In fact, the conditional expectation $E(u_{it+1}|t, z_{it}) = m(z_{it})$ where

z_{it} represents the vector of instruments. Consequently with idiosyncratic uninsurable risk and aggregate uncertainty the model is not identifiable using methods that rely on fixed T; we require methods that rely on large T asymptotics and in practice we need long enough time series of data that allow the aggregate shocks to average out. The suitable time series dimension depends on the variance of such shocks, with longer series required the higher the variance. However, we do not require to observe the same individual for a large number of time periods; just that the data covers long T.[39] Moreover, aggregate shocks cannot be accounted for using time dummies as emphasized by Altug and Miller (1990) unless there is no idiosyncratic uncertainty.

6.1.4.1. Linearizing the Euler equation A simpler way to go about estimation is to loglinearize (6.16):

$$-\Delta \ln \hat{v}'_{iyt+1} - \ln(1+r_t) = d_{it} + \ln\beta + \Delta \ln \psi'_{t+1} + \varepsilon_{it} \tag{6.17}$$

where

$$d_{it} = \ln\bigl[E_t\bigl\{\beta(1+r_t)\psi'_{t+1}\hat{v}'_{yt+1}\bigr\}\bigr] - E_t \ln\bigl[\beta(1+r_t)\psi'_{t+1}\hat{v}'_{yt+1}\bigr].$$

In the simplest case where the discounted marginal utility of consumption $mu_{it+1} = \beta(1+r_t)\psi'_{t+1}\hat{v}'_{yt+1}$ is a log-normal random variable, d_{it} will be proportional to its variance conditional on information in period t, i.e. $d_{it} = k \operatorname{Var}_t\{mu_{it+1}\}$. It is precisely this point that gives rise to the identification issue since the conditional variance will depend on variables relevant for predicting future income or wage realizations. However, if we are willing to restrict what the conditional variance depends on (and hence the stochastic process governing wages), this linearization offers a great simplification and often makes it easier to deal with measurement error in the underlying variables forming the marginal utility. Under nonnormality d_{it} will also depend on higher-order moments of the marginal utility of consumption mu_{it+1}.

Loglinearization has been widely used in the empirical analysis of consumption. However, identification in this case requires more restrictions than those implied by the theory. Its usage has been controversial [see Carroll (1997) and Ludvigson and Paxson (2001)] precisely because the basic exclusion restrictions used for identification in (6.16) may no longer be valid in (6.17). Implicitly linearization imposes restrictions on expectation formation and on the underlying process of uncertainty. Attanasio and Low (2002) examine these issues using Monte Carlo analysis in a wide variety of settings and conclude that in practice linearization is unlikely to bias the results in a serious way.

6.1.4.2. Accounting for corner solutions with no fixed costs When hours of work are at a corner solution the Euler condition (6.16) does not hold when evaluated at market prices. However, we can use the results of Heckman (1974a) and Neary and Roberts

[39] Meghir and Weber (1996) discuss this point in relation to estimating Euler equations.

(1980) to keep the Euler equation representation evaluated at shadow prices. Here we assume that there are no fixed costs of work and no search frictions and consequently that the participation decision is fully characterized by the standard reservation wage condition [Heckman (1974a)]. In particular nonworkers have a negative desired labor supply at the market wage corresponding to their skills, while workers have a positive desired labor supply, which is observed. It is easy to show that the intertemporal first-order conditions still hold, so long as we evaluate the indirect utility function at the shadow (reservation) wage w_{it}^R defined by

$$h\big(w_{it}^R, y_{it}\big) = 0. \tag{6.18}$$

Estimating the 'static' within-period labor supply function as described in earlier sections allows us to obtain a labor supply model that can then be solved for the reservation wage as in (6.18). In the next step the consumption Euler equation can be estimated using observed market wages for workers and shadow wages for nonworkers.[40]

6.1.4.3. An example Consider the labor supply model

$$h_{it} = \alpha(z_{it}) + \beta \ln w_{it} + \gamma \frac{y_{it}}{w_{it}} \tag{6.19}$$

where z_{it} are preference shifters such as household characteristics. This corresponds to a particular form of the indirect utility function presented in an earlier section. The term y is defined by $y = c - wh$, where w is the after tax wage and c is total household (non-durable) consumption, and hence is endogenous. The utility index can be computed by using the formula in (2.11). This gives the value of $\hat{v}_t$, from which $\hat{v}'_{yt}$ can be calculated. For workers the relevant wage will be the observed wage. For nonworkers the relevant wage at which to evaluate within-period utility is the reservation wage which is given by the positive solution for w in Equation (6.19) when $h = 0$, for given y. This has to be solved for numerically in this example. Using the reservation wage is equivalent to computing the direct utility function when hours are zero. This calculation is only valid if there are no fixed costs of work.

In the next step we can specify the part of the utility function that is not revealed by within-period choices. This is the monotonic transformation. One simple possibility would be to use a linear transformation; for example $\psi[v_t(w_{it}, y_{it})|z_{it}] = a(z_{it})v(c_{it}, h_{it})$, which would be interpretable as saying that characteristics z_{it} affect the discount rate. A more general alternative would be to allow characteristics to also affect the intertemporal substitution elasticity; for example $\psi[v_t(w_{it}, y)|z_{it}] = \frac{a(z_{it})}{1+\rho(z_{it})} v_{it}(w_{it}, y_{it})^{1+\rho(z_{it})}$, for some negative valued function $\rho(z_{it})$. The fact that all or some of the characteristics z affect within-period allocations does not imply that they will not also affect risk aversion or the way the future is discounted.

To obtain an example specification let $a(z_{it}) = 1$ and $\rho(z_{it}) = \rho_0 + \rho_1 fs_{it}$ where fs_{it} is family size for household i in period t. Using the utility function (2.11) term

[40] See Blundell, Meghir and Neves (1993).

$\hat{v}'_{yt} = (1+\hat{\gamma})^2 \frac{w^{\hat{\beta}}}{\hat{\beta}+1}$ can be evaluated at the estimated parameters. In this case the Euler equation for consumption over time will take the form

$$\hat{v}'_{iyt}\hat{v}_{it}^{\rho_0+\rho_1 fs_{it}} = E_t\{\beta(1+r_t)\hat{v}'_{iyt+1}\hat{v}_{it+1}^{\rho_0+\rho_1 fs_{it}}\}. \tag{6.20}$$

This can be estimated using nonlinear GMM treating the estimated marginal utility of money $\hat{v}'_{iyt}$ and the within-period utility index $\hat{v}_t$ as known [see Hansen (1982) and Hansen and Singleton (1982)]. The fact that the expression depends on estimated parameters does not affect consistency because as the sample size goes to infinity the parameters estimated on the first stage converge to the true values. Inference however requires us to correct the standard errors for the fact that we are relying on pre-estimated parameters.

The linearized version of the Euler equation here takes the form

$$\begin{aligned} &-\Delta \ln \hat{v}'_{iyt+1} - \ln(1+r_t) \\ &\quad = d_{it} + \ln\beta + \rho_o \Delta \ln \hat{v}_{it+1} + \rho'_1 \Delta \ln fs_{it}\hat{v}_{it+1} + \varepsilon_{it} \end{aligned}$$

which, given the assumptions implied by the loglinearization, can be estimated by linear GMM.

6.1.4.4. Testing for liquidity constraints One key issue for the interpretation of intertemporal behavior is the extent to which individuals are liquidity constrained which is defined as being able to borrow and save freely at a constant interest rate. It has been observed from very early on that consumption seems to track income, which is a fact often cited as evidence for liquidity constraints. However, this phenomenon can be explained within the model we have presented.

First, Heckman (1974b) has argued that such income tracking can be induced by nonseparability of consumption and labor supply: If consumption and leisure are strong enough substitutes, higher amounts of consumption will be related to higher levels of labor supply and hence higher income.

Second, family size and demographics, which affect consumption and labor supply allocations, evolve very much alongside income over the life-cycle, with family size growing when income grows most and declining when income declines [probably endogenously: see Blundell, Browning and Meghir (1994)]. By allowing for this in our model we have effectively accounted for another reason for tracking.

Finally, the evolution of the conditional variance of the marginal utility d_{it} also leads to consumption growth. This variance is likely to decline over the life-cycle as uncertainty is revealed. This is particularly true if shocks to wages are permanent or highly persistent. Thus a high d_{it} when young and a lower d_{it} when old will imply rapid consumption growth early on declining later, much like the evolution of income over the life-cycle [Carroll and Samwick (1998), Attanasio et al. (1999)].

The empirical challenge is to find sources of predictable income growth not already included in the model to account for preferences (e.g. nonseparability) and to test the hypothesis that they do not affect consumption growth. Browning and Collado (2001)

use the powerful idea of predictable changes in income due to pre-announced and regular seasonal bonuses in Spain and establish that consumption growth is not sensitive to these totally predictable changes in income. However we are not always as fortunate as that and we need to use other perhaps less compelling sources of predictable growth. One possibility is to include labor income growth. This is a useful source of variation for two reasons: Conditional on the wage rate, labor income would have variability because hours of work may change in a predictable way for other exogenous reasons. Second, hours should not enter the Euler equation once we also include wages. Nevertheless it is still an issue of what the exogenous source of hours would be that has not to do with preferences or changes in wages. Another possibility is to use predictable changes in other income. The problem is that income from investments, etc. is likely to be positive only for the wealthier individuals who are unlikely to want to borrow anyway.

Tests of liquidity constraints find no evidence of their importance once nonseparabilities and demographics are allowed for. This should not be interpreted as saying that anyone can borrow any amount they wish at a fixed rate; after all, the lack of complete markets is now generally accepted with moral hazard as its most probable source. However it may well mean that the lack of perfect credit markets is not important because individuals do not wish to borrow much against future income growth anyway when they would most need it (i.e. when young) because of uncertainty.

6.2. *Further issues in the specification and estimation of dynamic models of labor supply and consumption*

The model we have presented up to now in the context of intertemporal optimization lacks a number of potentially important features. These include unobserved preference heterogeneity, fixed costs of work and nonseparability over time. We now discuss these issues in turn and we complete our chapter by presenting the estimation of a model containing potentially all these features.

6.2.1. *Unobserved heterogeneity*

Allowing for unobserved preference heterogeneity seems like a natural step in constructing realistic models. Thus, for example, both MaCurdy (1981) and Heckman and MaCurdy (1980) recognize this and include fixed effects in their models. They recognize that preference heterogeneity could be persistent and may well be correlated with wages. The question is how to account for unobserved heterogeneity in a model without complete markets. The key difficulty stems from the fact that it is not possible to specify a model where both the Euler equation and the within-period condition have additive errors without restricting the structure of intertemporal preferences. Inevitably a model with unrestricted intertemporal preferences and unobserved heterogeneity will be nonseparable in unobservables. Standard orthogonality conditions do not suffice for identification in this case. In the Heckman and MaCurdy study the errors are effectively nonseparable because of the corner solutions. However, the complete markets assumption meant that a fixed effects Tobit estimator worked well even with moderate T.

There is a developing literature on the identification and estimation of models with nonseparable errors and endogenous regressors [e.g. Florens et al. (2007), Imbens and Newey (2007), Blundell and Powell (2004)], which provide alternative identifying conditions in this case. Even if one is to impose these stronger assumptions there remains the problem of finding suitable instruments, which are an ingredient of all such methods. The problem is particularly acute if unobserved heterogeneity is serially correlated, since the instruments are likely to be predetermined decisions. These difficulties will lead us to an estimation method based on a complete solution of the dynamic programming model.

6.2.2. *Estimating the intertemporal substitution model with fixed costs of work*

Fixed costs of work or other nonconvexities in the budget constraint pose a very serious challenge to the empirical analysis, even within a static framework. In this context the labor supply function is discontinuous at low hourly wage rates. Moreover as Cogan (1981) pointed out, the standard reservation wage which sets labor supply to zero does not generate a participation condition. Generally the participation and hours margins are explained by different models, which could be the result of the existence of fixed costs of work or of search frictions. The separation between the intensive and extensive margins (hours of work) requires extra identifying assumptions.

Within an intertemporal context fixed costs pose additional difficulties for modeling the participation decision. This involves a comparison between the life-cycle utility of work and nonwork, which requires solving the life-cycle model conditional on the person working and conditional on the person not working. Such a solution allows one to evaluate the current and future welfare consequences of the two decisions.

In the presence of fixed costs we can follow two empirical strategies. The first is a partial one and seeks to estimate the subset of parameters that are identifiable if one keeps labor supply behavior fixed. As such it cannot be informative for policy questions whose answer relies on the quantification of the complete labor supply and consumption response. However, it offers a way of testing some aspects of the life-cycle model in a relatively general setting and may be a first step in a stepwise approach for identifying the complete set of preferences.

The second approach specifies a complete structural model of labor supply and participation and uses methods from dynamic discrete choice to estimate labor supply responses. Before moving to a discussion of the full solution approach we briefly outline the conditional approach.

6.2.3. *The conditional Euler equation for consumption*

Consider the definition of the indirect utility function within period, based on a vector of goods $\mathbf{q}_t$ and prices $\mathbf{p}_t$ conditional on labor supply behavior h_t

$$v_t = \psi\big[v(c_t|\mathbf{p}_t, h_t), h_t\big] = \max_{\mathbf{q}}\big\{\psi\big[u(\mathbf{q}_t|h_t), h_t\big] \,\big|\, \mathbf{p}_t'\mathbf{q}_t = c_t\big\}. \tag{6.21}$$

We can then base the analysis of the intertemporal allocations on the utility index $v_t = \psi[v(c_t|\mathbf{p}_t, h_t), h_t]$. As in the case of the joint labor supply and consumption model presented earlier, all parameters implicit in $v(c_t|\mathbf{p}_t, h_t)$ can be estimated using a conditional (on h_t) within-period demand system [see Browning and Meghir (1991)]. This will depend on h_t if and only if the goods $\mathbf{q}_t$ are nonseparable from h_t. Under weak separability h_t will not affect demands directly. However, the intertemporal allocations can still depend on h_t without this having any implications for the structure of the within-period marginal rate-of-substitution functions between goods. This point has been noted now in several papers, all of which have demonstrated its empirical importance.[41]

The estimation approach is broadly similar to the one described above so we do not go over it again in detail. Once the within-period demand system characterizing the conditional choice of $\mathbf{q}_t$ has been estimated, we can construct the utility index $v(c_t|\mathbf{p}_t, h_t)$. The Euler equation for c_t can then be used to estimate the parameters of the function ψ up to an explicitly additive function of h_t. In general, the Euler equation as well as the demand system will be a function of h_t. This can include both hours of work as a continuous variable and indicators of whether the person is working or not, or other functions of h_t that are considered relevant. The crucial point to recognize however is that labor supply is endogenous both for within-period and for intertemporal allocations. Thus estimation requires suitable instruments. One possibility is to use lags in labor supply for this purpose. In the absence of unobserved heterogeneity the approach is valid. However, if persistent preference shocks have been ignored this approach could lead to inconsistent parameter estimates.

The conditional Euler equation for consumption provides a very powerful vehicle for testing the life-cycle model in relation to consumption behavior and for estimating some of the parameters in a way that is robust to the specific model of labor supply. In principle, hours of work can be determined in a number of ways, which we do not have to specify, subject to the proviso that we can specify instruments that can "predict" labor supply. However, from a policy perspective, the conditional Euler equation for consumption is of limited interest because it does not provide the full set of parameters required to answer even a simple partial equilibrium question. Thus a complete analysis of intertemporal labor supply and consumption needs to address directly estimation of a model for the determination of hours of work.

6.2.4. *Intertemporal nonseparability*

A final issue is whether preferences should be taken as separable over time.[42] It is well documented that labor supply behavior is very persistent which may be interpreted

[41] Attanasio and Weber (1993, 1995), Blundell, Browning and Meghir (1994) and Meghir and Weber (1996) all strongly reject the hypothesis that intertemporal allocations do not depend directly on observed labor supply.

[42] For some studies that relax intertemporal separability see Shaw (1989), Hotz, Kydland and Sedlacek (1988), Meghir and Weber (1996), Eckstein and Wolpin (1989) and Altug and Miller (1998).

as being due to nonseparability, although the source of persistence could well be unobserved heterogeneity. Another source of nonseparability can be the structure of the intertemporal budget constraint since current behavior may affect eligibility for welfare programs. Finally, if wages depend on past work experience, current work affects future earning prospects, which also leads to intertemporal nonseparability. These issues are considered in the next section.

6.3. Dynamic discrete choice models and intertemporal nonseparability

To address many of the issues presented above in a coherent and unified way we need to consider a complete model of life-cycle labor supply and consumption. This can be very complex and demanding on data. Thus in our presentation we start with a simplified model along these lines which ignores the saving decision but offers a way forward on the issue of fixed costs and nonseparability. We subsequently build on this to present a more complete model that includes saving.

One of the first attempts to model the dynamics of participation decisions when choices are discrete is given by Eckstein and Wolpin (1989). Their model concerns the labor supply of women. Husband's income is taken as exogenous. The within-period utility function, which is nonseparable in consumption c_t and participation p_t, takes the form

$$U_t = c_t + a_1 p_t + a_2 c_t p_t + a_3 p_t K_{t-1} + \sum_{j=1}^{J} a_{4j} N_{tj} p_t + a_5 p_t S \tag{6.22}$$

where K_{t-1} is the number of periods worked in the past; depending on the sign of a_3 this may turn out to reinforce work habits or not. The law of motion of K_t is simply $K_t = K_{t-1} + p_t$. Finally, S represents years of schooling and N_{tj} represents the number of children in age group j. This utility function in itself gives rise to intertemporal dependencies since current participation affects future preferences and a forward-looking individual will take this into account when making participation decisions. Further dynamics are induced by the budget constraint. This takes the form

$$y_t^w p_t + y_t^h = c_t + \sum_{j=1}^{J} \kappa_j N_{tj} p_t + b p_t \tag{6.23}$$

where κ_j are costs relating to children in the jth age group, b is a fixed cost of work and y_t^h is husband's income, which is taken to be an exogenous stochastic process, affecting female utility only through total resources. The female wage, y_t^w, depends on past work decisions:

$$\ln y_t^w = \beta_1 + \beta_2 K_{t-1} + \beta_3 K_{t-1}^2 + \beta_4 S + \varepsilon_t \tag{6.24}$$

where ε_t is an independently and identically distributed normal shock to wages. Hence the implied dynamics in this model are quite intricate: Past work decisions produce

human capital and enhance earnings potential. This should lead to increases in participation. On the other hand, past work decisions change preferences, either dampening down or reinforcing the effects due to enhanced human capital.

At this stage the only source of stochastic variation is the iid shock to wages ε_t. This formulation has the undesirable feature that the minimum observed wage is a consistent estimator for the reservation wage; this is because preferences are homogeneous in the population. To overcome this problem Eckstein and Wolpin allow observed wages to be measured with error, which turns out to be particularly important empirically. Thus observed wages satisfy

$$\ln y_t^{w*} = \ln y_t^w + u_t. \tag{6.25}$$

Eckstein and Wolpin assume that u_t is normally distributed.

In such dynamic discrete choice models estimation is complicated by the fact that participation in this period confers benefit/costs in future periods. Thus the future impact of current choices needs to be computed explicitly in order to compute the probability of participation. Eckstein and Wolpin follow a maximum likelihood approach where the parameters of the participation decision, of wages and of the measurement error process are estimated simultaneously.

Their estimation approach can be described as follows: An individual participates if the utility from doing so is higher than the utility from not working. To illustrate the approach we simplify further their model by assuming additive separability between consumption and participation. In this case the husband's income will not affect female labor supply. For notational simplicity we also drop the schooling (S) and household composition terms (N_{tj}). In this simplified framework, utility when participating can be written as

$$\begin{aligned} V_t^{(1)} &= y_t^w + y_t^h - b + a_1 + a_3 K_{t-1} + \delta E_t V_{t+1}(K_{t-1}+1) \\ &= \exp\bigl(\beta_1 + \beta_2 K_{t-1} + \beta_3 K_{t-1}^2 + \varepsilon_t\bigr) \\ &\quad + y_t^h - b + a_1 + a_3 K_{t-1} + \delta E_t V_{t+1}(K_{t-1}+1)_t \end{aligned} \tag{6.26}$$

while the utility from nonparticipation is given by

$$V_t^{(0)} = y_t^h + \delta E_t V_{t+1}(K_{t-1})_t, \tag{6.27}$$

where δ is the personal discount factor. Note that when the woman participates in this period, human capital increases by one; it does not increase otherwise. This is what gives rise to the difference in the future values associated with the current actions. In the expressions above the expectation is taken over the uncertain realizations of ε_t (and of the husband's income). This expectation is conditional on information known in period t. However, since the shock is iid, conditional and unconditional expectations coincide.

A participation rule can be derived now from these two expressions written in terms of thresholds for the unobserved shock ε_t. Workers are individuals with wage shocks

such that[43]

$$\begin{aligned}\varepsilon_t \geqslant\ & \ln\big[b - a_1 - a_3 K_{t-1} + \delta\big(E_t V_{t+1}(K_{t-1}) - E_t V_{t+1}(K_{t-1}+1)\big)\big] \\ & - \big(\beta_1 + \beta_2 K_{t-1} + \beta_3 K_{t-1}^2\big) \quad \text{or} \\ \varepsilon_t \geqslant\ & \varepsilon_t^*(K_{t-1}). \end{aligned} \tag{6.28}$$

Given a distributional assumption on ε_t this leads to a probability of participation. Note, however that the expression in (6.28) depends on the future expected gain from working. Hence to estimate the model this gain needs to be computed. This is achieved by backwards induction.

For a given set of parameters of the utility function and the distribution of the unobservable ε_t the value of participation and nonparticipation is constructed in a terminal period, given all possible values of the state variables (in this case K). For each K we then compute $E_t V_T(K)_t = E[\max(V_T^{(1)}, V_T^{(0)})]$ where the expectation is over the realizations ε_T. Computing the value in period T is very simple since the problem is essentially static then.

The only way by which past decisions affect the future is through the state variable K. Hence the future gain from working this period when the current experience stock is K is simply $E_t V_T(K+1) - E_t V_T(K)$. Whether this is positive or negative will depend on the effect of an extra unit of human capital on wages and on preferences. Given the terminal value function we can now compute the values in period $T-1$ for all possible K accumulated by period $T-1$ and so on until we reach period t. This computation is a simple recursion. The procedure requires one to specify a terminal period (age) T. It also requires us to be specific about what happens beyond that period. In models that require backwards induction it is often necessary to parameterize separately a terminal value function. In Eckstein and Wolpin the value beyond the last decision period T is assumed to be zero.

Given a way to compute $E_t V_{t+1}(K_{t-1}) - E_t V_{t+1}(K_{t-1}+1)$ we can now easily construct the likelihood function. For nonworkers this is simply $\Pr(\varepsilon_t < \varepsilon_t^*(K_{t-1})) = \Phi(\varepsilon_t^*(K_{t-1}))$ where Φ is the standard normal distribution function. For workers the contribution to the likelihood function is the joint density of wages (driven by the sum of the shock ε_t and the measurement error u_t) and the probability that $\varepsilon_t > \varepsilon_t^*(K_{t-1})$. Hence estimation proceeds as follows: For an initial set of parameters the future gains from work are computed. Then the observed event is computed and the likelihood function is constructed for each observation. A Newton-type algorithm can then be used to update the parameters. The value functions need to be recomputed at each iteration when updated parameters are available – this is what makes dynamic discrete choice computationally burdensome.

[43] In our simplified model the husband's income plays no role in the wife's decision. This is not a feature of the Eckstein and Wolpin model but a result of our simplified exposition in which we have assumed additive separability.

Estimation of this model requires observations on K_{t-1} and the choice p_t as well as wages. In general retrospective information on periods worked can be used, although entire work histories constructed over time as events unfold would reduce the chance of measurement error. Administrative data has now become available which improves the data situation substantially [see Adda et al. (2006)].

The dynamic discrete choice model described above is a coherent and powerful way of modeling the dynamics of participation and the evolution in wages. However, it does not allow for unobserved heterogeneity and thus all dependence on the past is in effect assumed to be pure state dependence.

The model by Eckstein and Wolpin is a prototype on which other researchers have built, drawing also from the experience gained in the analysis of discrete choice in other fields or in labor supply [Rust (1987), Pakes (1986), Hotz and Miller (1988), Keane and Wolpin (1997)]. One of the most important subsequent contributions in the field of labor supply is the paper by Rust and Phelan (1997). The crucial aspect of this paper is that it models explicitly the relationship between work and future social security entitlements, thus building a model that can be used to evaluate the impact of policy reforms. An important feature, which complicates the model and makes it much harder to implement, is that the individual's choice depends on a large number of state variables that evolve stochastically. In the Eckstein and Wolpin prototypical model there was basically only one state variable: the number of periods worked in the past. Here the state space includes health status, own earnings, spouse's earnings and social security income. Some of these variables are affected by past decisions. Hence the intertemporal nonseparabilities in this model are primarily induced by the structure of the budget set: Current work decisions affect both future earnings and future social security receipts.

The principle of estimating such a model does not differ fundamentally from that of estimating the Eckstein and Wolpin model: The stochastic process for the exogenous state variables is estimated from the data. Then, following the specification of a distribution for the unobservables, the probability of observed choices is constructed, which depends on the future and current utility gains from this choice. As before, for each set of parameter values and at each value of the state variables the model has to be solved and the optimal choice determined. The probabilities at each data point are combined in the usual way to form the sample likelihood function. However, the problem is more complicated because of the many sources of uncertainty, originating from the large number of stochastically evolving state variables. These components are critical additions because they recognize explicitly that there are events such as the possibility of death or taste shifter variables such as health that affect behavior but are fundamentally uncertain. Such uncertainty is very likely to affect labor supply and retirement behavior of individuals.

6.4. Estimation with saving, participation and unobserved heterogeneity

We conclude our chapter by a brief discussion of estimation of dynamic models with saving in the absence of complete markets, which brings together the entire set of issues

we have identified as challenges in estimating labor supply models and takes us right against the research frontier in this field.

6.4.1. Estimation with complete markets

Altug and Miller (1998) specify a model of consumption and labor supply, where preferences are nonseparable over time and where wages depend on past labor supply (experience). In a departure from the earlier literature, saving is explicitly taken into account as are aggregate shocks. Moreover, the estimation methods proposed are relatively simple since they exploit a modified version of the conditional choice probability estimator developed in Hotz and Miller (1993). The key assumption that allows them to estimate such a complex model is that markets are complete. They also assume that preferences for leisure and consumption are additive. Finally the problem is simplified further by assuming that preference shocks are independently and identically distributed over time (and individuals) and there is no source of persistent heterogeneity in preferences.

The complete markets assumption allows them to express consumption allocations as a function of a fixed effect and an aggregate time effect. This solves at one go the problem of dealing with aggregate shocks when the time period is short [Chamberlain (1984)] and the problem of having to simulate alternative consumption paths explicitly when solving the dynamic programing problem.

In Altug and Miller the complete markets assumption can be viewed as an approximation that allows them to estimate a more general economic model than the ones considered earlier in the literature. Indeed their model is particularly rich, because it allows for endogenous human capital accumulation and for nonseparable preferences as well as saving. However, the complete markets assumption is resoundingly rejected whenever it is tested [Cochrane (1991) and Attanasio and Davis (1996)]. It is not known how much bias the assumption would introduce in the parameter estimates. Nevertheless, the real empirical challenge is to relax both the complete markets assumption and the structure of unobserved heterogeneity. In the next section we review the issues surrounding this challenge.

6.4.2. Estimation with uninsurable idiosyncratic risk

We consider an economy where some idiosyncratic risk remains uninsurable. However we assume that perfect credit markets are available.[44] Consider a utility function depending on hours of work h_{it}, on participation p_{it} (to reflect fixed costs) and on consumption c_{it}:

$$U_{it} = U_1(c_{it}, h_{it}, p_{it}, f_i|z_{it}) + U_2(h_{it}, f_i|z_{it}) + \gamma(z_{it})p_{it} + p_{it}v_{it}^{(1)} + (1 - p_{it})v_{it}^{(0)} \tag{6.29}$$

[44] Some may view this as a contradiction. However, given uncertainty, most individuals will typically not want much uncolateralized borrowing, making the modeling of liquidity constraints probably redundant for all practical purposes. This may be why many tests for liquidity constraints fail to reject the null of no constraints.

where z_{it} are taste shifter variables and where f_i, $\nu_{it}^{(1)}$ and $\nu_{it}^{(0)}$ are heterogeneity terms, the first being time invariant. Assets accumulate according to the difference equation

$$A_{it+1} = (1 + r_t)(A_{it} + w_{it}h_{it} - c_{it}).$$

The terminal condition for assets is

$$A_{iT} = 0,$$

where T is the last period of the planning horizon. We do not discuss retirement explicitly. However, early retirement can be induced by the availability of pensions later in life, by the accumulation of private assets, by aspects of the welfare system such as easily available disability insurance and/or by a decline in wages at an older age.

We assume wages take the form

$$\ln w_{it} = d_t^e + \kappa_i + \zeta^{ed'} x_{it} + e_{it}$$

where d_t^e is the log price of human capital for education group e, x_{it} denotes observable characteristics, some of which may be common with z_{it}, κ_i is a fixed effect and e_{it} is an iid shock with a known distribution, say normal.[45]

Suppose the function U_1 in (6.29) is nonadditive in participation p, hours h and consumption c with no components that are additive in p or h. In this case it is possible to estimate U_1 and U_2 based on the conditional Euler equation for consumption and on the within-period labor supply decision as discussed earlier, subject to being able to deal with unobserved heterogeneity. However, the function γ cannot be identified in this way. This missing component will be key to simulating counterfactual employment, hours and consumption paths for individuals. Despite the relative simplicity of preferences and the wage function, both of which exclude intertemporal dependencies, the estimation of all relevant parameters requires the full solution of the dynamic optimization problem: the probability of working is a function of the utility gain from doing so. To compute this utility gain one must know the consumption in the counterfactual state. With incomplete markets and idiosyncratic shocks this is not as straightforward as in the Altug and Miller case. We outline a possible approach.

We start by simplifying the model and assume a constant interest rate $r_t = r$. Next specify the conditional distribution governing the evolution of all other state variables, i.e. $g_s(S_{it}|S_{it-1}, \ldots, S_{it-p})$, where S includes all stochastically time-varying characteristics in x and z taken to be exogenous. In general g_s can be estimated separately and we can condition on it during estimation of the rest of the model.

In general heterogeneity in the wage rate κ_i will be correlated with the heterogeneity in preferences f_i. This implies that wages are endogenous for both labor supply and consumption and this reflects the idea that unobserved productivity and the tastes for work are related. A simplifying assumption could be made reducing the dimension of heterogeneity, e.g. $f_i \propto \kappa_i$.

[45] Richer stochastic structures are in principle possible, but they do increase the state space substantially.

In this model assets are the only endogenous state variable, which in principle should include all sources of household wealth, including housing and pension wealth. This causes a very serious measurement problem. Leaving this aside, given suitable data the model is solved numerically to obtain the value of consumption conditional on the person's labor market state. Denote the optimal solutions as follows: workers $c_{it}^{(1)} = c_t^{(1)}(w_{it}, A_{it}|S_{it}, f_i, p_{it} = 1)$, nonworkers $c_{it}^{(0)} = c_t^{(0)}(A_{it}|S_{it}, f_i, p_{it} = 0)$ and $h_{it}^{(1)} = h_t(w_{it}, A_{it}|S_{it}, f_i, p_{it} = 1)$. In general there will be no closed form solutions to these functions and they will need to be computed numerically during estimation. To compute these policy functions we need to solve for the future optimal policies. One approach for this finite horizon problem is to use backwards induction. Starting from some terminal period, the optimal policies are evaluated for all possible values of the state variables backwards up until the current period. At this point we have all the ingredients to evaluate the probability of work, including $c^{(1)}$ and $c^{(0)}$ and the future values conditional on current actions working ($EV_{it+1}^{(1)}$) and not working ($EV_{it+1}^{(0)}$). The current value of working and not working are then given by

$$V_{it}^{(1)} = U\big(c_{it}^{(1)}, h_{it}^{(1)}, p_{it} = 1, f_i\big) + \nu_{it}^{(1)} + \beta E_t V_{it+1}^{(1)},$$
$$V_{it}^{(0)} = U\big(c_{it}^{(0)}, h_{it} = 0, p_{it} = 0, f_i\big) + \nu_{it}^{(0)} + \beta E_t V_{it+1}^{(0)},$$

which now allows us to specify the probability of working as

$$\begin{aligned}&\Pr(p_{it} = 1|A_{it}, S_{it}, f_i)\\ &\quad = \Pr\big(\nu_{it}^{(1)} - \nu_{it}^{(0)} > U_{it}^{(0)} - U_{it}^{(1)} + \beta\big[E_t V_{it+1}^{(0)} - E_t V_{it+1}^{(1)}\big]\big).\end{aligned}$$

The consumption and labor supply as derived above are deterministic given the fixed effect f_i. The reason for this is that the time-varying heterogeneity terms $v^{(1)}$ and $v^{(0)}$ do not affect the marginal utility of hours (given participation) or consumption. One simple way to enrich the stochastic specification is to allow for measurement error in consumption and hours. This will induce a density of observed hours m_h among workers and observed consumption $m_c^{(1)}$ for workers and $m_c^{(0)}$ for nonworkers. Thus the likelihood conditional on the heterogeneity term is

$$\begin{aligned}\mathcal{L} = \prod_{i=1}^{N}\prod_{t=1}^{T_i}\Bigg\{&\big[m_h m_c g(w_{it}|\kappa_i, S_{it})\Pr(p_{it} = 1|w_{it}, A_{it}, S_{it}, f_i)\big]^{p_{it}}\\ &\cdot\left[\int\big[m_c^{(0)} g(w_{it}|f_i, S_{it})\big(1 - \Pr(p_{it} = 1|w_{it}, A_{it}, S_{it}, f_i)\big)\big]\,dw_{it}\right]^{1-p_{it}}\Bigg\}\\ &\cdot\prod_{i=1}^{N}\prod_{t=1}^{T_i}\mathcal{L}_{it}(f_i)\end{aligned}$$

where $g(\cdot)$ is the density of wages, N is the number of individuals, T_i is the number of time periods over which individual i is observed and $\mathcal{L}_{it}(f_i)$ is the likelihood contribution for individual i. The stochastic dependence between the various elements in

the likelihood is driven by the unobserved component f_i, which needs to be integrated out.

Allowing for persistent unobserved heterogeneity is complicated by the fact that at any point in time f_i will be correlated with assets: these are the outcome of past decisions, themselves a function of f_i. Thus in a panel of individual data the initial value of assets cannot be taken as exogenous in general. To solve this problem we need to specify a model for the initial value (A_{i0}), conditional on a set of variables assumed themselves to be exogenous. Denote the distribution of initial assets by $g_A(A_{i0}|\zeta_i, z_{it})$ where z_{it} are a set of instruments explaining initial assets, which are excludable from the participation probability. Finding such instruments is not straightforward. One possibility could be to use random shocks that affected wealth at some point, but did not change preferences, such as, for example, parental health. The unobserved variables ζ_i and f_i may be correlated, which is the source of endogeneity of initial assets. If these are exogenous, f_i and ζ_i would be independent of each other.[46]

Given a model for initial assets and using a discrete mixture as an approximation to the distribution of the pair (f_i, ζ_i) [see Heckman and Singer (1984)] the likelihood function now becomes

$$\mathcal{L} = \prod_{i=1}^{N} \sum_{k=1}^{K} \sum_{s=1}^{S} \left\{ \mathrm{pr}_{ks}\, g_A(A_{i0}|\zeta_s, z_{it}) \prod_{t=1}^{T_i} \mathcal{L}_{it}(f_k) \right\}$$

where K and S are the number of points of support for the distribution of f_i and ζ_i respectively and pr_{ks} is the probability mass at a point of the (f_i, ζ_i) distribution.

The computational burden in these models arises from having to solve the model at each iteration and each individual type (defined by the observable and unobservable characteristics) for all values of the state variables. If these are continuous (such as assets) they need to be discretized.

Macroeconomic shocks The model allows for macroeconomic shocks through wages. In its simplest form there is just one type of human capital and the time effect on the wage reflects its value relative to the consumption good. In a richer setting there are different types of human capital with relative prices that vary. To allow for macro-shocks in the model we require a model that predicts forward prices as a function of current observables. In principle, this process will have to be estimated simultaneously with the model, because of the changes in labor force composition over time, which the model accounts for.

6.4.3. *Why allow for saving?*

Allowing for saving is complicated both computationally and empirically. Allowing for a linear utility in consumption would eliminate the complications. So why should we get

[46] See Ham and LaLonde (1996) and Meghir and Whitehouse (1997) for applications in dynamic transition models.

into all this trouble? The answer lies in the fact that individuals are risk averse and risk is not fully insurable. Modeling saving in this context is important for understanding a number of issues, including self-insurance for events such unemployment [Low (1999)] and more importantly pensions and retirement. For example, to understand the policy impact of changes in pension arrangements we need to understand how such policies interact with saving. The extent to which public policies crowd out private saving can only be studied in a model that accounts for both. Similar issues will arise when studying the impact of policies such as taxes and tax credits. The complete labor supply effect cannot be understood if we do not know how saving behavior will be affected. On the other hand, there are many questions relating to whether our fully rational forward-looking model is a good enough representation of reality. Ignoring the issue is, however, not the way forward.

7. Summary and conclusions

The study of labor supply is valuable from a number of perspectives. The analysis of the impact of taxes and benefits is perhaps the best-established motivation. Within this field we are concerned with the impact of taxes on effort as well as the role of taxes and benefits in affecting education decisions; in this latter case labor supply is seen as an alternative to school or training for younger individuals. From a more dynamic perspective, focus recently has also shifted to labor supply as a way of responding to uncertainty and mitigating the amount of saving as well as for understanding the evolution of consumption over the life-cycle: without allowing for changes in labor supply, it is very difficult to rationalize the observed behavior of consumption. Finally, the relationship of consumption and labor supply is critical for understanding issues to do with optimal taxes and the design of benefits – in-work benefits in particular. For all the above reasons, it is clearly important to understand the way labor supply is determined and how this relates to intertemporal considerations, such as saving.

This chapter outlines a number of approaches to the study of labor supply beginning with the original static models and ending with dynamic ones that allow for saving and possibly intertemporal nonseparabilities. Along the way we have discussed incorporating taxes and allowing for nonconvex budget sets and the importance of unobserved heterogeneity. Allowing for the last has proved particularly important empirically for estimating reliable models that are capable of fitting the data and accounting for the large persistence in labor supply patterns. Empirically, labor supply analysis poses significant challenges not only because of the nonconvexities but also because of the endogeneity of the main variables whose effect we are attempting to measure. High-effort people are likely to have invested more in human capital and thus have higher wages. They also accumulate more wealth, making asset income potentially endogenous as well. Adding dynamics and allowing for nonconvexities in the budget sets compounds the difficulties. We have attempted to provide a flavor of these difficulties and point to solutions. However, it is clear that there is more to be done. One relatively new and important area

of research which we did not touch upon is modeling the entire career, starting with education choice and continuing with labor supply over the life-cycle. This is likely to be of key importance for understanding the longer-term impact of public policy: programs, such as tax credits, that encourage labor supply may well discourage education. Trading off these two margins of adjustment is important and requires reliable models for both. Thus, considering the dynamics of labor supply and developing reliable modeling methods will continue to be of key importance for policy purposes.

Appendix A

This appendix reviews general formulations for likelihood functions applicable to econometric models involving any combination of five types of endogenous variables: (1) discrete, (2) continuous, (3) censored, (4) truncated, and (5) continuous-discrete. The subsequent discussion opens with an overview of the statistical framework considered here. It next considers increasingly complex variants of this framework, starting with models incorporating just discrete variables, adding in continuous variables, and then including endogenous variables of a combined continuous-discrete character. The analysis proceeds to cover specifications appropriate when one does not observe all states of the world but instead only knows whether various combinations of states have occurred. The concluding subsection presents alternative representations of likelihood functions commonly found in the literature comparable to the specifications presented here, as well as presenting simple extensions of specifications that allow for dependence on exogenous variables.

A.1. Overview of statistical framework

The basic idea at the foundation of econometric models characterizing distributions of discrete-continuous variables relies on the notion that all endogenous quantities depend on the values of an underlying set of continuously-distributed random variables. Specify these underlying variables by the vector U, assumed to include r linearly-independent components. This $r \times 1$ vector possesses the joint density function

$$\varphi(U) \quad \text{for } U \in \Omega \tag{A.1}$$

where the set Ω designates the sample space or domain of the random variables U.

In this model, m states of the world can occur. The discrete random variable δ_i signifies whether state i happens, with $\delta_i = 1$ indicating realization of state i and $\delta_i = 0$ implying that some state other than i occurred. The value of δ_i depends on where U falls in its sample space; specifically,

$$\delta_i = \begin{cases} 1 & \text{if } U \in \Omega_i, \\ 0 & \text{otherwise,} \end{cases} \tag{A.2}$$

where the set Ω_i represents a nontrivial subset of the entire sample space Ω. Without loss of generality, assume that the sets Ω_i for $i = 1, \ldots, m$ are mutually exclusive and exhaustive, meaning $\bigcup_{i=1}^{m} \Omega_i = \Omega$ and the sets $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$ (i.e., the sets Ω_i and Ω_j are disjoint). In association with state i, there exist n_i continuously distributed random variables designated Y_{ji}, $j = 1, \ldots, n_i$. The following equations determine the values of these continuous variables:

$$Y_{ji} = g_{ji}(U). \tag{A.3}$$

Stacking these individual random variables into a vector yields

$$Y_i = \begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{n_i i} \end{pmatrix} = \begin{pmatrix} g_{1i} \\ \vdots \\ g_{n_i i} \end{pmatrix} = g_i. \tag{A.4}$$

To avoid introduction of redundant or ill-defined Y_{ji}'s, assume there exists an inverse of g_i such that

$$U_{(i)} = g_i^{-1}(Y_i, U_i) \tag{A.5}$$

for some subvector $U_{(i)}$ comprised of any n_i components of U.[47] The subvector U_i includes those elements of U not included in $U_{(i)}$. Designate Φ_i as the domain of (Y_i, U_i) and Θ_i as the domain of Y_i.

Another interesting class of random variables consists of quantities that take a fixed single value in some states and a continuous set of values in others. Denote these discrete-continuous variables as Z_{ji}, with the index i signaling the state realized and $j = 1, \ldots, k_i$ signifying the particular Z realized in this state. The value of Z_{ji} follows a rule of the form

$$Z_{ji} = \begin{cases} Y_{ji} & \text{for } j \in K_{ci}, \\ Z_{ji}^* & \text{for } j \in K_{di}, \end{cases} \tag{A.6}$$

where the set K_{ci} indexes those Z_{ji} taking the form of a continuous variable in state i, and the set K_{di} identifies those Z_{ji} equaling a constant value Z_{ji}^* in state i. Define Z_i as the vector containing the Z_{ji}, $j = 1, \ldots, k_i$, as elements analogous to Y_i specified in (A.4).

Finally, form all the unique variables appearing in any of the Y_i's into the vector Y, assumed to be of dimension $n \times 1$, and all the variables making up the Z_i's into the vector Z, assumed to be of dimension $k \times 1$. For any event $\delta_i = 1$, Y consists of two sets of components: the vector Y_i incorporating all the continuous random variables registering in state i, and $Y_{(i)}$ made up of all other continuous variables unobserved in this state but

[47] Assuming existence of the inverse of g_i in (A.5) is not as restrictive as one might first surmise. If an inverse does not exist on set Φ_i, then one can replace Φ_i with a further segment of this set with inverses defined on each of these smaller sets. The subsequent analysis can then be carried out for this expanded decomposition of Φ.

seen in some other state $j \neq i$. Similarly, Z consists of the vector Z_i and $Z_{(i)}$ defined analogously. In some states i, all of the elements of Y and Z may be observed, and in others none may be. The subsequent discussion characterizes formulations of conditional and unconditional likelihood functions associated with Y, Z and combinations of the δ_i's. As briefly noted at the end of this appendix, one can readily introduce the presence of exogenous variables both in specifying the distribution of U and in defining the regions of definition of δ_i. An exogenous variable in this analysis must be observed in all states; otherwise, this variable must be included as a component of Y or Z.

A.2. *Discrete variables: All and combinations of states*

Initially consider empirical frameworks in which one observes only discrete variables whose outcomes register the realization of m distinct regimes determined by the relative values of U.

A common formulation specifies that a researcher sees exactly which state i occurs, implying that one observes all individual δ_i, $i = 1, \ldots, m$. From (A.2) we see that the probability that $\delta_i = 1$ equals

$$P(\delta_i = 1) = P(U \in \Omega_i) = \int \cdots \int_{\Omega_i} \varphi(U)\, dU \equiv \int_{\Omega_i} \varphi(U)\, dU. \tag{A.7}$$

The notation $\int \cdots \int_{\Omega_i}$ denotes integration over the set Ω_i, which the end of this equation expresses in the shorthand notation $\int_{\Omega_i}$. The joint distribution of the δ_i's takes the form

$$P(\delta_1, \ldots, \delta_m) = \prod_{i=1}^{m} \big[P(\delta_i = 1)\big]^{\delta_i} = \prod_{i \in M} \big[P(\delta_i = 1)\big]^{\delta_i}. \tag{A.8}$$

In the last part of this equation, the notation $M = \{i\colon\ i = 1, \ldots, m\}$ refers to the set of all possible states i.

In other formulations, a researcher does not observe or chooses to ignore each state individually. Instead, one accounts for only whether some combination of states has been realized. More specifically, suppose one knows that at least one $\delta_i = 1$ when $i \in M_t \subset M$, but one does not account for which particular δ_i in this group actually occurred. So,

$$\text{if } i \in M_t, \text{ then } \overline{\delta}_t \equiv \sum_{i \in M_t} \delta_i = 1; \quad \text{otherwise, } \overline{\delta}_t = 0. \tag{A.9}$$

The sets M_t, $t = 1, \ldots, \tau$, are mutually exclusive and exhaustive (i.e., $\bigcup_{t=1}^{\tau} M_t = M$ and $M_t \cap M_j = \emptyset$ for $t \neq j$). The probability of the occurrence of group state t equals

$$P(\overline{\delta}_t = 1) = \sum_{i \in M_t} P(\delta_i = 1). \tag{A.10}$$

The joint distribution of the $\bar{\delta}_t$'s takes the form

$$P(\bar{\delta}_1, \ldots, \bar{\delta}_\tau) = \prod_{t \in T} \left[P(\bar{\delta}_t = 1)\right]^{\bar{\delta}_t} \tag{A.11}$$

where the notation $T = \{t\colon\ t = 1, \ldots, \tau\}$ refers to the set of all possible group states t.

A.3. Continuous variables: All states observed

Consider those models in which one observes each individual δ_i along with vectors Y_i of continuously distributed random variables for states $i \in M_y \subseteq M$. Conditional on occurrence of a state, the components of Y_i may either be truncated or censored. The truncated elements of Y_i refer to those that lie in a strict subset of their overall domain given realization of the selection mechanism $U \in \Omega_i$ (or, equivalently, $(Y_i, U_i) \in \Phi_i$). The censored elements consist of those that instead range over their entire domain. The set $\Theta = \bigcup_{i=1}^m \Theta_i$ defines the sample space of Y. So, if Y_i includes truncated components, then $\Theta_i \subset \Theta$.

The first step in formulating specifications for the distributions of the Y_i's involves recognizing that the density of underlying random variables U conditional on the event $\delta_i = 1$ takes the form

$$\varphi(U \mid \delta_i = 1) = \frac{\varphi(U)}{P(\delta_i = 1)} \tag{A.12}$$

where relationship (A.7) gives the formula for $P(\delta_i = 1)$. An alternative expression (A.7) is given by

$$P(\delta_i = 1) = P(U \in \Omega_i) = P\big((Y_i, U_i) \in \Phi_i\big) = \int_{\Phi_i} h_i(Y_i, U_i)\, dY_i\, dU_i \tag{A.13}$$

where the set $\Phi = \bigcup_{i=1}^n \Phi_i$ defines the domain of $(Y, U_1, \ldots, U_n)$.

Application of a conventional change-in-variables formula exploiting relations (A.3) and (A.5) yields the following specification for the density of Y_i conditional on δ_i:

$$f(Y_i \mid \delta_i = 1) = \frac{\int_{\Phi_{i|Y_i}} h_i(Y_i, U_i)\, dU_i}{P(\delta_i = 1)} \quad \text{for } Y_i \in \Theta_i \tag{A.14}$$

where

$$h_i(Y_i, U_i) = J_i \varphi\big(g_i^{-1}(Y_i, U_i), U_i\big) \quad \text{with } J_i = \left|\frac{\partial g_i^{-1}}{\partial Y_i'}\right|^+, \tag{A.15}$$

and the notation $\int_{\Phi_{i|Y_i}}$ denotes integration of U_i over the set

$$\Phi_{i|Y_i} = \big\{U_i\colon\ (Y_i, U_i) \in \Phi_i\big\}. \tag{A.16}$$

The term J_i in (A.15) represents the Jacobian of the transformation associated with (A.5) (i.e., J_i is the absolute value of the determinant of the matrix of partial derivatives

$\frac{\partial g_i^{-1}}{\partial Y_i'}$). One can express the domain of Y_i as

$$\Theta_i = \Theta_{i \cdot Y_i} = \{Y_i \colon U_i \in \Phi_{i|Y_i}\}, \tag{A.17}$$

where the notation $\Theta_{i \cdot Y_i}$ simply signifies that this set is a subspace of Y_i.

A compact expression for the conditional density of Y is

$$f(Y \mid \delta_i,\ i \in M_y) = \prod_{i \in M_y} \big[f(Y_i \mid \delta_i = 1)\big]^{\delta_i}, \tag{A.18}$$

where as defined above M_y designates the set of states in which one observes at least one element of Y. An alternative representation for this conditional density takes the form

$$f(Y \mid \delta_1, \ldots, \delta_m) = \prod_{i \in M_y} \big[f(Y_i \mid \delta_i = 1)\big]^{\delta_i} \prod_{i \in M_y^c} [1]^{\delta_i}, \tag{A.19}$$

where the set M_y^c denotes the complement of M_y with respect to M. Realizations of $i \in M_y^c$ mean that all elements of Y are either undefined or unobserved.

The joint density of Y and $\delta_1, \ldots, \delta_m$ is the product of the conditional density of Y given by (A.19) and the joint probability of $\delta_1, \ldots, \delta_m$ given by (A.8), yielding

$$\begin{aligned} f(Y, \delta_1, \ldots, \delta_m) &= \prod_{i \in M_y} \big[f(Y_i \mid \delta_i = 1) P(\delta_i = 1)\big]^{\delta_i} \prod_{i \in M_y^c} \big[P(\delta_i = 1)\big]^{\delta_i} \\ &= \prod_{i \in M_y} \left[\int_{\Phi_{i|Y_i}} h_i(Y_i, U_i)\, dU_i\right]^{\delta_i} \prod_{i \in M_y^c} \left[\int_{\Omega_i} \varphi(U)\, dU\right]^{\delta_i}. \end{aligned} \tag{A.20}$$

The second line of this expression follows by substituting relationships from (A.7) and (A.14).

A.4. Discrete/continuous variables: All states observed

Consider models in which one observes individual δ_i along with the vectors Z_i comprised of discrete-continuous random variables for states $i \in M_z \subseteq M$. The components included in Z_i are either distributed continuously or equal to constants according to the following rule:

$$Z_i = \begin{pmatrix} Z_{ci} \\ Z_{di} \end{pmatrix} = \begin{pmatrix} Y_i \\ Z_{di}^* \end{pmatrix} \quad \text{for } i \in M_z. \tag{A.21}$$

Inspection of (A.6) reveals that those individual Z_{ji} for $j \in K_{ci}$ make up the elements of the vector Z_{ci}; and those Z_{ji} for $j \in K_{di}$ form the vector Z_{di}. The set M_z comprises all states in which any component of Z is realized.

For states $i \in M_y$, one can express the distribution of Z_i conditional on $\delta_i = 1$ as

$$f(Z_i \mid \delta_i = 1) = f(Z_{ci}, Z_{di} \mid \delta_i = 1)$$

$$= f(Z_{ci} \mid Z_{di}, \delta_i = 1) P(Z_{di} \mid \delta_i = 1)$$
$$= f(Y_i \mid Z^*_{di}, \delta_i = 1) \tag{A.22}$$

where the third line follows from

$$P(Z_{di} = Z^*_{di} \mid \delta_i = 1) = 1. \tag{A.23}$$

Formally, the argument Z^*_{di} in $f(Y_i \mid Z^*_{di}, \delta_i = 1)$ is redundant since the event $\delta_i = 1$ already implies $Z_{di} = Z^*_{di}$; the argument is included merely to remind the reader that the density appearing in the last line of (A.22) typically depends on Z^*_{di}. A compact expression for the conditional density of Z is

$$f(Z \mid \delta_i, i \in M_z) = \prod_{i \in M_y} \left[f(Y_i \mid Z^*_{di}, \delta_i = 1) \right]^{\delta_i} \prod_{i \in M_d} [1]^{\delta_i}. \tag{A.24}$$

Realizations of $i \in M_y$ mean that some of the elements of Z_i are continuously distributed, whereas occurrence of $i \in M_d$ implies that all elements of Z_i are discrete. One can write an alternative representation for this conditional density as

$$\begin{aligned} f(Z \mid \delta_1, \ldots, \delta_m) &= \prod_{i \in M_y} \left[f\big(Y_i \mid Z^*_{di}, \delta_i = 1\big) \right]^{\delta_i} \prod_{i \in M_d} [1]^{\delta_i} \prod_{i \in M_u} [1]^{\delta_i} \\ &= \prod_{i \in M_y} \left[f\big(Y_i \mid Z^*_{di}, \delta_i = 1\big) \right]^{\delta_i} \prod_{i \in M_d \cup M_u} [1]^{\delta_i}. \end{aligned} \tag{A.25}$$

Realizations of $i \in M_u$ mean that all components of Z are either undefined or unknown.

The joint density of Z and $\delta_1, \ldots, \delta_m$ is the product of the conditional density of Z given by (A.25) and the joint probability of $\delta_1, \ldots, \delta_m$ given by (A.8), yielding

$$\begin{aligned} f(Z, \delta_1, \ldots, \delta_m) &= \prod_{i \in M_y} \left[f\big(Y_i \mid Z^*_{di}, \delta_i = 1\big) P(\delta_i = 1) \right]^{\delta_i} \prod_{i \in M_d \cup M_u} \left[P(\delta_i = 1) \right]^{\delta_i} \\ &= \prod_{i \in M_y} \left[\int_{\Phi_{i|Y_i, Z^*_{di}}} h_i(Y_i, U_i)\, dU_i \right]^{\delta_i} \prod_{i \in M_d \cup M_u} \left[\int_{\Omega_i} \varphi(U)\, dU \right]^{\delta_i}. \end{aligned} \tag{A.26}$$

The second line of this expression follows by substituting relationships from (A.7) and (A.14), where the notation $\Phi_{i|Y_i, Z^*_{di}}$ still refers to the set $\Phi_{i|Y_i}$ defined by (A.16) with emphasis added to indicate that this set also depends on Z^*_{di}.

A.5. Discrete/continuous variables: Combinations of states

An important category of models involves characterizing the distribution of continuous and discrete-continuous variables when one either observes or chooses to distinguish the occurrence of groups rather than individual states. Define the relevant groups of states by the $\bar{\delta}_t$'s specified in (A.9) for $t \in T$ as outlined in Section A.2.

Consider the distribution of the continuous random variable

$$Y_t = \sum_{i \in M_t} \delta_i Y_i. \tag{A.27}$$

Relation (A.27) implicitly assumes that each Y_i is defined and of comparable dimension for $i \in M_t$. Application of the law of iterated expectations yields the following density for Y_t conditional on $\bar{\delta}_t = 1$:

$$\begin{aligned} f(Y_t \mid \bar{\delta}_t = 1) &= \sum_{i \in M_t} f(Y_t \mid \delta_i = 1, \bar{\delta}_t = 1) P(\delta_i = 1 \mid \bar{\delta}_t = 1) \\ &= \sum_{i \in M_t} f(Y_t \mid \delta_i = 1) P(\delta_i = 1 \mid \bar{\delta}_t = 1) \\ &= \sum_{i \in M_t} f(Y_t \mid \delta_i = 1) \frac{P(\delta_i = 1)}{P(\bar{\delta}_t = 1)}. \end{aligned} \tag{A.28}$$

The latter two lines of this relationship follow from the assumptions that the individual states $\delta_i = 1$ for $i \in M_t$ making up the event $\bar{\delta}_t = 1$ are mutually exhaustive and exclusive.

Discrete-continuous variables are realized according to the following rule:

$$Z_t = \begin{pmatrix} Z_{ct} \\ Z_{dt} \end{pmatrix} = \begin{pmatrix} Y_t \\ Z_{dt}^* \end{pmatrix} \quad \text{for } t \in T_z. \tag{A.29}$$

The set $T_z = T_y \cup T_d$ comprises all group states in which any component of Z is realized. The set T_y includes those group states t in which Z_t incorporates the continuously-distributed vector Y_t specified by (A.27); and the set T_d includes those group states wherein all the components of Z_t equal constant values.[48]

For group states $t \in T_c$, the distribution of Z_t conditional on $\bar{\delta}_t = 1$ takes the form

$$\begin{aligned} f(Z_t \mid \bar{\delta}_t = 1) &= f(Z_{ct}, Z_{dt} \mid \bar{\delta}_t = 1) \\ &= f(Z_{ct} \mid Z_{dt}, \bar{\delta}_t = 1) P(Z_{dt} \mid \bar{\delta}_t = 1) \\ &= f\big(Y_i \mid Z_{dt}^*, \bar{\delta}_t = 1\big), \end{aligned} \tag{A.30}$$

where this latter expression exploits the relationship

$$P\big(Z_{dt} = Z_{dt}^* \mid \bar{\delta}_t = 1\big) = 1 \quad \text{for } t \in T_d. \tag{A.31}$$

Analogous to (A.25), a compact expression for the conditional density of Z is

$$f(Z \mid \bar{\delta}_1, \ldots, \bar{\delta}_\tau) = \prod_{t \in T_y} \big[f\big(Y_t \mid Z_{dt}^*, \bar{\delta}_t = 1\big) \big]^{\bar{\delta}_t} \prod_{t \in T_d \cup T_u} [1]^{\bar{\delta}_t} \tag{A.32}$$

[48] For notational simplicity, the specification of the values of Z_t when $t \in T_d$ presumes that Z_t^* is common across the individual states $i \in M_t$ making up group state t. One can instead replace the common value Z_t^* by a set $\{Z_t^*\}$ consisting of several discrete values at the expense of introducing some complexity in specifying likelihood functions.

where the set T_u includes those state groups in which no Z_{jt} are either undefined or unknown.

Multiplying the conditional density (A.32) by the joint probability of the events $\bar{\delta}_1, \ldots, \bar{\delta}_\tau$ given by (A.8) generates the following joint density for Z and the $\bar{\delta}_t$'s:

$$
\begin{aligned}
&f(Z, \bar{\delta}_1, \ldots, \bar{\delta}_\tau) \\
&\quad = \prod_{t \in T_y} \left[f\left(Y_t \mid Z^*_{dt}, \bar{\delta}_t = 1\right) P(\bar{\delta}_t = 1) \right]^{\bar{\delta}_t} \prod_{t \in T_d \cup T_u} \left[P(\bar{\delta}_t = 1) \right]^{\bar{\delta}_t} \\
&\quad = \prod_{t \in T_y} \left[\sum_{i \in M_t} \int_{\Phi_{i|Y_t Z^*_{dt}}} h_i(Y_t, U_i)\, dU_i \right]^{\bar{\delta}_t} \prod_{t \in T_d \cup T_u} \left[\sum_{i \in M_t} \int_{\Omega_i} \varphi(U)\, dU \right]^{\bar{\delta}_t}. \qquad \text{(A.33)}
\end{aligned}
$$

The last line of this expression follows from substitution of relationships from (A.7), (A.10), (A.14), and (A.28).

A.6. Accounting for unobserved and exogenous variables

Specification (A.33) presents a general formulation for likelihood functions incorporating discrete, continuous and discrete-continuous variables. One often sees alternative representations of this specification in the literature that may at first not appear as a special case of (A.33).

One such representation defines a set of continuous or discrete-continuous variables Z that are then presumed to be unobserved and, therefore, must be eliminated as arguments of the f's in (A.33). In particular, suppose Z consists of two components $Z'_t = (Z'_{1t}, Z'_{2t})$ where the variables Z'_{1t} are observed and those included in Z'_{2t} are not. Correspondingly, decompose $Y'_t = (Y'_{1t}, Y'_{2t})$ and $Z'_{dt} = (Z'_{1dt}, Z'_{2dt})$, with the random variables Y'_{2t} and Z'_{2dt} unobserved.

Integrating (or summing) the joint likelihood function (A.33) over Z'_{2t} produces the marginal distribution for Z'_{1t}. This exercise yields

$$
\begin{aligned}
&f(Z_1, \bar{\delta}_1, \ldots, \bar{\delta}_\tau) \\
&\quad = \prod_{t \in T_y} \int_{\Theta_{t \cdot\, Y_{2t}}} f\left(Y_{1t}, Y_{2t} \mid Z^*_{dt}, \bar{\delta}_t = 1\right) dY_{2t}\, P(\bar{\delta}_t = 1)^{\bar{\delta}_t} \prod_{t \in T_d \cup T_u} \left[P(\bar{\delta}_t = 1) \right]^{\bar{\delta}_t} \\
&\quad = \prod_{t \in T_y} f\left(Y_{1t} \mid Z^*_{d1t}, \bar{\delta}_t = 1\right) P(\bar{\delta}_t = 1)^{\bar{\delta}_t} \prod_{t \in T_d \cup T_u} \left[P(\bar{\delta}_t = 1) \right]^{\bar{\delta}_t}. \qquad \text{(A.34)}
\end{aligned}
$$

The last line of this expression exploits the relationship

$$
\begin{aligned}
&\int_{\Theta_{t \cdot\, Y_{2t}}} f\left(Y_{1t}, Y_{2t} \mid Z^*_{dt}, \bar{\delta}_t = 1\right) dY_{2t} \\
&\quad = \int_{\Theta_{t \cdot\, Y_{2t}}} f\left(Y_{2t} \mid Y_{1t}, Z^*_{dt}, \bar{\delta}_t = 1\right) dY_{2t}\, f\left(Y_{1t} \mid Z^*_{dt}, \bar{\delta}_t = 1\right)
\end{aligned}
$$

$$= f\big(Y_{1t} \mid Z_{dt}^*, \overline{\delta}_t = 1\big),$$

which follows since $\Theta_{t\cdot\, Y_{2t}}$ constitutes the domain of Y_{2t} given the event $\overline{\delta}_t = 1$. Clearly, the last line in (A.34) is a special case of (A.33). This merely reflects the fact that an unobserved Y_{2t} has been reinterpreted as a component of the U_i's implicit in (A.33). The variables making up U_i in a state i (or t) may be observed as a Y_j in some other state.

Finally, throughout the above discussion one can readily interpret the distribution of U as being conditional on a set of exogenous variables X, as well as define the regions of definition of δ_i to depend on X (so, $\Omega_i = \Omega_i(X)$). To be deemed exogenous, each component of X must be observed in all states; otherwise, this variable must be treated as a component of Y or Z in the previous analysis. Modifying the above formulae to admit exogenous X merely involves adding X as an argument of $f(\cdot)$ and interpreting the sample subspaces Ω_i, $\Phi_{i|Y_i}$, and Θ_i as functions of X.

References

Adda, J., Dustmann, C., Meghir, C., Robin, J.M. (2006). "Career progression and formal versus on-the-job training". IFS Working Paper W06/16.

Altonji, J.G. (1982). "The intertemporal substitution model of labor market fluctuations: An empirical analysis". Review of Economic Studies 49, 783–824.

Altonji, J.G. (1986). "Intertemporal substitution in labor supply: Evidence from micro data". Journal of Political Economy 94 (June, Part II), S176–S215.

Altug, S., Miller, R. (1990). "Household choices in equilibrium". Econometrica 58 (May), 543–570.

Altug, S., Miller, R. (1998). "The effect of work experience on female wages and labor supply". Review of Economic Studies 65 (January), 45–85.

Arellano, M., Meghir, C. (1992). "Female labor supply and on-the-job search: An empirical model estimated using complementary data sets". Review of Economic Studies 59(3) (200), 537–559.

Arrufat, J.L., Zabalza, A. (1986). "Female labor supply with taxation, random preferences, and optimization errors". Econometrica 54, 47–63.

Attanasio, O., Banks, J., Meghir, C., Weber, G. (1999). "Humps and bumps in lifetime consumption". Journal of Business and Economic Statistics 17 (1), 22–35.

Attanasio, O., Davis, S. (1996). "Relative wage movements and the distribution of consumption". The Journal of Political Economy 104 (6, December), 1227–1262.

Attanasio, O., Low, H. (2002). "Estimating Euler equations". IFS Working Paper WP02/06.

Attanasio, O.P., Weber, G. (1993). "Consumption growth, the interest rate and aggregation". Review of Economic Studies 60, 631–649.

Attanasio, O.P., Weber, G. (1995). "Is consumption growth consistent with intertemporal optimization? Evidence from the Consumer Expenditure Survey". The Journal of Political Economy 103 (6, December), 1121–1157.

Blackorby, C., Primont, D., Russell, R. (1978). Duality, Separability, and Functional Structure: Theory and Economic Applications. North-Holland, Amsterdam.

Blomquist, N.S. (1983). "The effect of income taxation on the labor supply of married men in Sweden". Journal of Public Economics 22, 169–197.

Blomquist, N.S. (1996). "Estimation methods for male labor supply functions: How to take account of nonlinear taxes". Journal of Econometrics 70, 383–405.

Blomquist, N.S., Hansson-Brusewitz, U. (1990). "The effect of taxes on male and female labor supply in Sweden". Journal of Human Resources 25, 317–357.

Blomquist, N.S., Newey, W. (2002). "Nonparametric estimation with nonlinear Budget Sets". Econometrica 70 (6), 2455–2480 (November).
Blundell, R.W., Browning, M., Meghir, C. (1994). "Consumer demand and the life-cycle allocation of household expenditure". Review of Economic Studies 161, 57–80.
Blundell, R.W., Chiappori, P.-A., Magnac, T., Meghir, C. (2007). "Collective labor supply and participation". Review of Economic Studies 74 (2), 417–445 (April).
Blundell, R.W., Chiappori, P.-A., Meghir, C. (2005). "Collective labor supply with children". Journal of Political Economy 113 (6, December), 1277–1306.
Blundell, R.W., Duncan, A., Meghir, C. (1992). "Taxation and empirical labor supply models: Lone parents in the UK". Economic Journal 102, 265–278.
Blundell, R.W., Duncan, A., Meghir, C. (1998). "Estimating labor supply responses using tax policy reforms". Econometrica 66, 827–861.
Blundell, R.W., MaCurdy, T. (1999). "Labor supply: A review of alternative approaches". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics. North-Holland, Amsterdam.
Blundell, R.W., Meghir, C., Neves, P. (1993). "Labour supply: An intertemporal substitution". Journal of Econometrics 59, 137–160.
Blundell, R.W., Powell, J. (2004). "Endogeneity in semiparametric binary response models". Review of Economic Studies 71 (5), 665–679.
Blundell, R.W., Walker, I. (1986). "A life cycle consistent empirical model of labor supply using cross section data". Review of Economic Studies 53, 539–558.
Bound, J., Jaeger, D.A., Baker, R.M. (1995). "Problems with instrumental variables estimation when the correlation between the instrument and the endogenous explanatory variable is weak". Journal of the American Statistical Association 90, 443–450.
Browning, M., Chiappori, P.-A. (1998). "Efficient intra-household allocations: A general characterization and empirical tests". Econometrica 66 (6), 1241–1278 (November).
Browning, M., Collado, M.D. (2001). "The response of expenditures to anticipated income changes: Panel data estimates". The American Economic Review 91 (3), 681–692 (June).
Browning, M., Deaton, A., Irish, M. (1985). "A profitable approach to labor supply and commodity demand over the life cycle". Econometrica 53 (May), 503–543.
Browning, M., Meghir, C. (1991). "Testing for separability between goods and leisure using conditional demand systems". Econometrica 59 (July), 925–952.
Carroll, C.D. (1997). "Death to the log-linearized Euler equation! (and very poor health to the second order approximation)". NBER Working Paper 6298.
Carroll, C.D., Samwick, A.A. (1998). "How important is precautionary saving?". Review of Economics and Statistics 80 (3), 410–419.
Chamberlain, G. (1984). "Panel data". In: Griliches, Z., Intriligator, M. (Eds.), Handbook of Econometrics, vol. 2. North-Holland, Amsterdam. Chapter 22.
Chiappori, P.-A. (1988). "Rational household labor supply". Econometrica 56, 63–90.
Chiappori, P.-A. (1992). "Collective labor supply and welfare". Journal of Political Economy 100, 437–467.
Cochrane, J. (1991). "A simple test of consumption insurance". The Journal of Political Economy 99 (5, October), 957–976.
Cogan, J. (1980). "Labor supply with costs of labor market entry". In: Smith, J. (Ed.), Female Labor Supply: Theory and Estimation. Princeton University Press, Princeton, pp. 327–364.
Cogan, J.F. (1981). "Fixed costs and labor supply". Econometrica 49, 945–964.
Darolles, S., Florens, J.P., Renault, E. (2000). "Nonparametric instrumental regression". Mimeo.
Deaton, A. (1974). "A reconsideration of the empirical implications of additive preferences". The Economic Journal 84 (334), 338–348 (June).
Eckstein, Z., Wolpin, K. (1989). "Dynamic labor force participation of married women and endogenous work experience". Review of Economic Studies 56, 375–390.
Eissa, N., Liebman, J. (1995). "Labor supply responses to the earned income tax credit". NBER Working Paper no. 5158.

Florens, J.P., Heckman, J.J., Meghir, C., Vytlacil, E. (2007). "Identification of treatment effects using control functions in models with continuous endogenous treatment and heterogeneous effects". IFS WP 07.

Fortin, B., Lacroix, G. (1997). "A test of neoclassical and collective models of household labor supply". Economic Journal 107, 933–955.

Gorman, W.M. (1959). "Separable utility and aggregation". Econometrica 21, 63–80.

Gorman, W.M. (1968). "The structure of utility functions". The Review of Economic Studies 35, 369–390.

Gosling, A., Machin, S., Meghir, C. (2000). "The changing distribution of male wages in the UK". Review of Economic Studies 67, 635–666.

Gronau, R. (1974). "Wage comparisons, a selectivity bias". Journal of Political Economy 82 (6), 1119–1144.

Ham, J.C., LaLonde, R.J. (1996). "The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training". Econometrica 64 (1), 175–205 (January).

Hansen, L.P. (1982). "Large sample properties of generalized method of moments estimators". Econometrica 50 (4), 1029–1054 (July).

Hansen, L.P., Singleton, K.J. (1982). "Generalized instrumental variables estimation of nonlinear rational expectations models". Econometrica 50 (5), 1269–1286 (September).

Härdle, W., Linton, O. (1994). Applied Nonparametric Methods. In: Handbook of Econometrics. North-Holland. Chapter 4.

Hausman, J. (1980). "The effect of wages, taxes and fixed costs on women's labor force participation". Journal of Public Economics 14, 161–194.

Hausman, J. (1981). "Labor supply". In: Aaron, H., Pechman, J. (Eds.), How Taxes Affect Economic Behavior. Brookings Institution, Washington, DC.

Hausman, J. (1985a). "The econometrics of nonlinear budget sets". Econometrica 53, 1255–1282.

Hausman, J. (1985b). "Taxes and labor supply". In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 1. North-Holland, Amsterdam.

Heckman, J.J. (1974a). "Shadow prices, market wages and labor supply". Econometrica 42, 679–694.

Heckman, J.J. (1974b). "Life-cycle consumption and labor supply: An explanation of the relationship between income and consumption over the life cycle". American Economic Review 64 (March), 188–194.

Heckman, J.J. (1974c). "Effects of child-care programs on women's work effort". Journal of Political Economy 82 (2), S136–S163.

Heckman, J.J. (1976). "Life-cycle model of earnings, learning and consumption". Journal of Political Economy 84, S11–S44.

Heckman, J.J. (1979). "Sample selection bias as a specification error". Econometrica 47, 153–161.

Heckman, J.J. (1993). "What has been learned about labor supply in the past twenty years?". American Economic Review (Papers and Proceedings) 83, 116–121.

Heckman, J.J., MaCurdy, T.E. (1980). "A life-cycle model of female labor supply". Review of Economic Studies 47, 47–74.

Heckman, J.J., Robb, R. (1985). "Alternative methods for evaluating the impact of interventions". In: Heckman, J.J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. In: Econometric Society Monograph, vol. 10. Cambridge University Press.

Heckman, J.J., Singer, B. (1984). "A method for minimizing the impact of distributional assumptions in econometric models for duration data". Econometrica 52 (2), 271–320 (March).

Honore, B.E. (1992). "Trimmed lad and least squares estimation of truncated and censored regression models with fixed effects". Econometrica 60 (3, May), 533–565.

Hotz, V.J., Miller, R.A. (1988). "An empirical analysis of life cycle fertility and female labor supply". Econometrica 56 (1), 91–118 (January).

Hotz, V.J., Miller, R.A. (1993). "Conditional choice probabilities and the estimation of dynamic models". The Review of Economic Studies 60 (3), 497–529 (Jully).

Hotz, V.J., Kydland, F.E., Sedlacek, G.L. (1988). "Intertemporal substitution and labor supply". Econometrica 56 (March), 335–360.

Hoynes, H.W. (1996). "Welfare transfers in two-parent famileis: Labor supply and welfare participation under AFDC-UP". Econometrica 64 (2), 295–332.

Imbens, G., Angrist, J. (1994). "Identification and estimation of local average treatment effects (in notes and comments)". Econometrica 62 (2, March), 467–475.
Imbens, G., Newey, W. (2007). "Identification and estimation in trianglar simultaneous equations models without additivity". April MIT mimeo.
Keane, M.P., Moffitt, R. (1998). "A structural model of multiple welfare program participation and labor supply". International Economic Review 39, 553–589.
Keane, M.P., Wolpin, K.I. (1997). "The career decisions of young men". The Journal of Political Economy 105 (3, June), 473–522.
Killingsworth, M. (1983). Labor Supply. Cambridge University Press, Cambridge.
Killingsworth, M., Heckman, J. (1986). "Female labor supply: A survey". In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, vol. 1. North-Holland, Amsterdam, pp. 103–204.
Klein, R.W., Spady, R.H. (1993). "An efficient semiparametric estimator for binary response models". Econometrica 61, 387–421 (March).
Kyriazidou, E. (1997). "Estimation of a panel data sample selection model". Econometrica 65 (6), 1335–1364 (November).
Lee, L.F. (1984). "Tests for the bivariate normal distribution in econometric models with selectivity". Econometrica 52 (4), 843–864.
Low, H. (1999). "Self-insurance and unemployment benefit in a life-cycle model of labor supply and savings". IFS Working Paper 99/24.
Ludvigson, S., Paxson, C. (2001). "Approximation bias in linearized Euler equations". Review of Economics and Statistics 83 (2, May), 242–256.
MaCurdy, T.E. (1981). "An empirical model of labor supply in a life-cycle setting". Journal of Political Economy 89, 1059–1085.
MaCurdy, T.E. (1983). "A simple scheme for estimating an intertemporal model of labor supply and consumption in the presence of taxes and uncertainty". International Economic Review 24, 265–289.
MaCurdy, T.E. (1985). "Interpreting empirical models of labour supply in an intertemporal framework with uncertainty". In: Heckman, J.J., Singer, B. (Eds.), Longitudinal Analysis of Labour Market Data. Cambridge University Press, Cambridge. Chapter 3.
MaCurdy, T.E. (1992). "Work disincentive effects of taxes: A re-examination of some evidence". American Economic Review 82, 243–249.
MaCurdy, T.E., Green, D., Paarsch, H. (1990). "Assessing empirical approaches for analyzing taxes and labor supply". Journal of Human Resources 25, 415–490.
Manski, C., McFadden, D. (1981). Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA.
Meghir, C., Weber, G. (1996). "Intertemporal nonseparability or borrowing restrictions: A disaggregate analysis using a US consumption panel". Econometrica 64, 1151–1181.
Meghir, C., Whitehouse, E. (1997). "Labour market transitions and retirement of men in the UK". Journal of Econometrics 79, 327–354.
Moffitt, R. (1983). "An economic model of welfare stigma". American Economic Review 73, 1023–1035.
Moffitt, R. (1986). "The econometrics of piecewise-linear budget constraints: Survey and exposition of the maximum likelihood method". Journal of Business and Economic Statistics 4, 317–327.
Mroz, T.A. (1987). "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions". Econometrica 55, 765–800.
Neary, J.P., Roberts, K.W.S. (1980). "The theory of household behaviour under rationing". European Economic Review 13, 25–42.
Newey, W.K., Powell, J.L. (2003). "Instrumental variable estimation of nonparametric models". Econometrica 71 (5), 1568–1578.
Newey, W.K., Powell, J.L., Vella, F. (1999). "Nonparametric estimation of triangular simultaneous equations models". Econometrica 67, 565–604.
Newey, W.K., Powell, J.L., Walker, J.R. (1990). "Semiparametric estimation of selection models: Some empirical results". American Economic Review 80 (2), 324–328.

Pakes, A. (1986). "Patents as options: Some estimates of the value of holding European patent stocks". Econometrica 54 (4), 755–784 (July).

Pencavel, J. (1986). "Labor supply of men: A survey". In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, vol. 1. North-Holland, Amsterdam, pp. 3–102.

Powell, J.L. (1987). "Semiparametric estimation of bivariate latent variable models". WP 8704SSRI, University of Wisconsin-Madison, July.

Robinson, P. (1988). "Semiparametric econometrics: A servey". Journal of Applied Econometricy 35, 35–51.

Rust, J. (1987). "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher". Econometrica 55 (5), 999–1033 (September).

Rust, J., Phelan, C. (1997). "How social security and medicare affect retirement in a world of incomplete markets". Econometrica 65 (4), 781–831.

Shaw, K. (1989). "Life-cycle labor supply with human capital accumulation". International Economic Review 30 (2), 431–457.

Thomas, D. (1990). "Intra-household resource allocation: An inferential approach". The Journal of Human Resources 25 (4), 635–664 (Autumn).

AUTHOR INDEX OF VOLUMES 6A AND 6B

n indicates citation in a footnote.

SUBJECT INDEX OF VOLUMES 6A AND 6B