FOUNDATIONS OF DEMAND ESTIMATION

By

Steven T. Berry and Philip A. Haile

September 2021

COWLES FOUNDATION DISCUSSION PAPER NO. 2301

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Foundations of Demand Estimation[*]

Steven T. Berry
Yale University

Philip A. Haile
Yale University

September 16, 2021

Abstract

Demand elasticities and other features of demand are critical determinants of the answers to most positive and normative questions about market power or the functioning of markets in practice. As a result, reliable demand estimation is an essential input to many types of research in Industrial Organization and other fields of economics. This chapter presents a discussion of some foundational issues in demand estimation. We focus on the distinctive challenges of demand estimation and strategies one can use to overcome them. We cover core models, alternative data settings, common estimation approaches, the role and choice of instruments, and nonparametric identification.

[*] For the *Handbook of Industrial Organization.* We benefited from detailed comments from Chris Conlon, Jean-Pierre Dubé, Aviv Nevo, and Frank Verboven. Jaewon Lee provided capable research assistance.

Contents

1 Introduction

1.1 Why Estimate Demand?

Little can be said about the functioning of a market without a quantitative assessment of demand. In the modern Industrial Organization ("IO") literature, for example, estimation of demand elasticities is essential for measuring markups and quantifying sources of market power. And almost any counterfactual question about a market requires quantitative measures of how choices respond to *ceteris paribus* changes in the prices or other characteristics of the available options. Such measures seldom suffice on their own to provide full answers to the economic questions of interest concerning market outcomes, but they are often necessary. Examples include assessments of

- the impact of a tax or subsidy;
- the social value of a new good;
- the effect of a tariff on prices;
- the effect of a merger on consumer prices;
- outcomes under alternative school choice regimes;
- the impact of adverse selection on insurance markets; and
- any quantitative question (and many qualitative questions) concerning consumer welfare.

Of course, the importance of measuring demand is not limited to the realm of IO. As these examples suggest, demand is central to substantive policy questions in public economics, trade, health, education, and other fields.

The problem of demand estimation is not new. However, it is both challenging and subject to some common misconceptions. It has also been the focus of substantial attention in recent years. Critical contributions have come from scholars in a range of fields. But IO economists often have led the way to the recent progress in demand estimation. One reason is the essential role that demand plays in questions of competition, market power, and market outcomes—questions at the heart of the field. The tradition in modern IO has also been to insist on trying to measure what is important to its core questions, rather than focusing on what is easy to measure given current data and estimation techniques. Determining what to measure and how to do so often requires guidance from an economic model, particularly in the context of imperfect competition and market equilibrium. This has led to a strong connection between theory and empirical work in IO. All of this has forced IO economists to face some difficult issues associated with demand estimation and to press for new solutions.

1.2 Our Focus

Our goal in this chapter is to provide a unified and (reasonably) compact treatment of some key ideas and practices developed in several literatures. Our emphasis naturally reflects our own perspectives. We limit our focus to a set of foundational issues: the distinctive challenges of demand estimation; the most common empirical models of demand in IO; the role and choice of instrumental variables; different data types (market-level data, micro data, panel data); and nonparametric identification. This focus implies that we neglect many other important issues, variations of the standard models, and applications. The chapter of Gandhi and Nevo in this *Handbook* has a different and highly complementary focus.

We begin with a discussion of the special challenges posed by the problem of demand estimation. We then review the classic discrete choice model of demand and move on to estimation and identification using market-level data. We next turn to various forms of "micro" (consumer-level) data and the advantages such data can offer. We conclude with a discussion of directions for future research.

2 The Challenges of Demand Estimation

Demand estimation is surprisingly difficult. At the heart of the challenge is the need to measure responses of quantities demanded to *ceteris paribus* changes in prices or other factors. The most basic requirements for credible estimation of such effects are standard in empirical economics; for example, one needs sufficiently flexible functional forms, valid sources of exogenous variation, and sufficient account for unobserved individual-level heterogeneity. But some distinctive challenges of demand estimation appear when one acknowledges the presence of unobservables that affect demand even when aggregating to the level of the product and market. The effects of such unobserved "demand shocks" arise through elementary economics and lead to the well-recognized problem of price endogeneity. Critically, the relevant demand shocks generally must be held fixed to measure the demand elasticities and other demand responses of economic interest. This alone rules out some common empirical approaches to estimation with endogeneity, even in the hypothetical case of a single-good economy. And the challenges become more severe in the presence of complements or substitutes with their own prices and demand shocks—i.e., in most real applications. In such settings, each good's demand depends on multiple endogenous prices and multiple demand shocks. As we discuss below, even when one's interest is limited to demand for a single good, these demand shocks introduce complications that are outside the realm of everyday empirical economics.

2.1 The First Fundamental Challenge

A primary challenge in demand estimation is the endogeneity of prices; i.e., statistical dependence between prices and unobservables that also affect demand. To illustrate, imagine a perfectly competitive market with a single good (no complements or substitutes), where demand and supply are characterized by the pair of simultaneous equations

$$\begin{aligned} Q &= D(X, P, U) && (2.1) \\ P &= C(W, Q, V). && (2.2) \end{aligned}$$

Here (2.1) represents demand while (2.2) represents supply. Q and P denote quantity and price; X and W denote observable demand shifters and cost shifters, respectively; and the "error terms" $U \in \mathbb{R}$ and $V \in \mathbb{R}$ denote unobserved demand shifters and cost shifters (shocks to demand and marginal cost), respectively. Assume that only prices and quantities are endogenous, i.e., that $(X, W) \perp\!\!\!\perp (U, V)$.

The presence of the latent demand shifters and cost shifters (U, V) is, of course, the reason there is a simultaneity/endogeneity "problem" in the identification/estimation of demand.[1] If there were no unobservable U affecting demand, identification would be trivial: the directly observed relationship between (X, P) and Q would itself be the demand relationship. In that case, the notion that the price P could be econometrically endogenous would be nonsense: there would be no unobservable whose variation could be confounded with that of P.

Of course, even when one acknowledges the existence of latent demand shocks U, the solution to the endogeneity problem in this special setting is well understood. Identification of demand can then be obtained through cost shifters that are excluded from the function D (i.e., elements of $W \not\subset X$) satisfying an appropriate independence condition with respect to U. In fact, nonparametric identification of demand in this case follows immediately from the same instrumental variables ("IV") conditions that yield identification in the case of a regression model with endogeneity.[2]

There is at least one important sense in which the optimistic message from this special case is correct: as we will see below, the essential requirements for identification of demand are indeed standard types of "clean" variation, as from instrumental variables.[3] However, this special case can also be highly misleading. It may suggest that the challenge to identification of demand is "merely" the endogeneity of price. And given the wide

[1] Unless marginal cost (inverse supply) is constant with respect to Q, dependence between P and U would arise even if U and V were assumed independent. Typically one will expect correlation between latent factors affecting firm costs (V) and those affecting consumer demand (U). And outside the realm of perfect competition, prices will generally depend on demand elasticities, yielding (functional) dependence between prices and demand shocks (the latter affecting elasticities) even in the special case of constant marginal cost and independence between cost shocks and demand shocks.

[2] See Newey and Powell (2003) in the case of nonparametric D that is additively separable in U, or Chernozhukov and Hansen (2005) (Theorem 4) when additive separability is replaced with strict monotonicity in U.

[3] Related to the instrumental variables literature is the literature on restrictions on the matrix of covariances across equations. For example, one might be willing in some contexts to assume that the demand and supply unobservables are uncorrelated. Full or partial identification using such restrictions dates back Koopmans (1949). A recent example in the oligopoly context is MacKay and Miller (2021).

attention to endogeneity in empirical economics, one might conclude that estimation of demand could proceed using any number of common tools for measuring the (causal) effects of endogenous covariates on outcomes of interest. As we explain below, both of these conclusions would be incorrect.

2.2 The Second Fundamental Challenge

A second (and less familiar) fundamental challenge is the fact that demand for any one good generally depends on more than one latent demand shock.

Demand for a given good, of course, typically depends on the prices and characteristics of all related goods. Indeed, when students first learn the elementary supply and demand model, they are taught that demand for one good cannot be considered in isolation. For example, a change in the price (or quality) of a substitute or complement will cause demand to shift. But in this regard there is nothing special about the observability (to us) of the prices or characteristics of related goods: demand for the good of interest also shifts in response to variation in the unobserved demand shifters associated with other goods. This is true regardless of whether supply is assumed to be perfectly competitive.

Thus, elementary economics tells us that the demand for any one good will in general depend on the prices, observed characteristics, and demand shocks *of that good and all related goods*. None of these factors can, in general, be excluded from the demand for a given good. Such factors are, therefore, among the "all else" that is held fixed when defining a *ceteris paribus* effect, such as that of a change in one price. This creates further challenges for demand estimation. It implies that many familiar econometric tools cannot be used to estimate demand unless one is willing to rely on strong functional form assumptions for identification. For example, demand estimation typically cannot be treated as standard regression analysis.

2.3 Demand Is Not Regression

Suppose there are J interrelated goods in a market of interest. Then demand for each good $j = 1, \ldots, J$ takes the form

$$Q_j = D_j \left(X, P, U\right), \tag{2.3}$$

where now $X = \left(X_1, \ldots, X_J\right), P = \left(P_1, \ldots, P_J\right), U = \left(U_1, \ldots, U_J\right)$.

Notice that J structural errors—the demand shocks $(U_1, \ldots, U_J)$ associated with all goods—enter on the right-hand side of (2.3). In general, this is not a regression equation.[4] For example, having valid instruments for all J prices will not generally suffice for identification of D_j or the *ceteris paribus* effects of price changes. We will see this

[4] A regression equation is most often specified in the separable form $Y = f(X) + E$ (e.g., Newey and Powell (2003)), leading to mean regression. Alternatively, a quantile regression model takes the form $Y = f(X, E)$, with f strictly increasing in the scalar E (e.g., Chernozhukov and Hansen (2005).)

formally in later sections of this chapter; but it is clear that results for identification of a regression function with a scalar structural error (see, e.g., footnote 2) are not directly applicable.

In general, econometric models with multiple structural errors in each equation are much more challenging than regression models. However, such models are not foreign to econometrics. For example, they arise in standard models of treatment effects (e.g., Angrist and Imbens (1995)) and in particular representations—typically the reduced forms—of standard simultaneous equations models (e.g., Brown (1983) and Benkard and Berry (2006)). These examples, in fact, hint at both the problems created by multiple structural errors and how one might make progress.

In empirical settings with endogeneity and multiple unobservables, economists often settle for estimation of particular weighted average responses (e.g., a local average treatment effect); but this is a compromise poorly suited to the economic questions that motive demand estimation, as these typically require the levels and slopes of demand at specific points. To make progress, the IO literature on demand estimation has used tools more familiar in the literature on simultaneous equations models: first "inverting" the system of equations, then relying on instrumental variables. Of course, in the case of simultaneous equations models,[5] one typically expects to need instruments (or other sources of exogenous variation) for all endogenous variables: here, all J prices and all J quantities. Indeed, although this is sometimes not fully appreciated, the IO literature has developed strategies exploiting sources of independent variation in prices and quantities even when estimating demand alone. We return to each of these points below.

2.4 A Surprisingly Difficult Case: Exogenous Prices

To appreciate the distinctive challenge created by the structural errors on the right-hand side of (2.3) it is useful to consider a hypothetical experimental setting. Suppose a researcher is able to randomly assign price vectors $(p_{1t}, \ldots, p_{Jt})$ in many large markets t and observe the resulting quantities $(q_{1t}, \ldots, q_{Jt})$ demanded in each market.[6] It may be natural to imagine that identification of demand would be trivial in this case. It is not.

The problem is that the economic quantities of interest—demand elasticities being the leading case—require that prices be varied while *holding all else fixed*. Among the things that must be held fixed are the demand shocks. Assigning prices for each market at random can avoid dependence between the demand shocks and prices, but it will not hold the demand shocks fixed.

The observed variation in quantities with the randomized variation in prices will re-

[5] See, e.g., Matzkin (2008, 2015), Blundell, Kristensen, and Matzkin (2017), and Berry and Haile (2018).

[6] Note that we assume the quantity observed is the quantity demanded rather than the quantity supplied at the exogenously set price. This need not be the case when a price is imposed exogenously—say, by random variation around a regulatory threshold—but could arise, e.g., through a marketing experiment.

veal certain averages of demand responses—integrating over the vector of demand shocks $(U_1, \ldots, U_J)$ to reveal a type of local average treatment effect. But in general, such averages are of very limited value in the case of demand, as they do not reveal any elasticity of demand (or other standard notion of a demand response)—not at the observed prices and quantities or any other known point. They therefore do not allow one to quantify demand responses to counterfactuals of interest (e.g., outcomes following a merger), to predict pass-through of a tax, or to infer firm markups through equilibrium pricing conditions. We return to this issue in section 2.5.3. In short, however, such averages offer at most a descriptive feature of demand, not the primary quantities of economic interest.

Of course, if experimental variation in prices does not allow identification of demand, it should be clear that "quasi-experimental" variation cannot suffice. One important implication is that instruments for prices (or other quasi-experimental variation in prices) generally cannot by themselves deliver identification of demand. This point is often not fully appreciated, but is essential to motivating the strategies relied on in the leading work on demand estimation.

One possible path to resolving this challenge is to impose functional form restrictions. For example, suppose that the demand function on the right-hand side of (2.3) is restricted to take the form

$$D_j\left(X, P, U\right) = D_j\left(X, P, \mathcal{E}_j(U)\right) \tag{2.4}$$

where the index $\mathcal{E}_j(U)$ is a scalar and the function D_j is strictly increasing in $\mathcal{E}_j(U)$. This structure would arise, for example, if demand for good j is linear in each of the demand shocks $U_1, \ldots, U_J$. In this case, as in Matzkin (2003), the τ quantile of the distribution of $Q_j|X, P$ reveals $D_j\left(X, P, \mathcal{E}_j(U)\right)$ for $\mathcal{E}_j(U)$ fixed at its τ quantile. Thus, in this special case, identification of demand can indeed be obtained when $P \perp\!\!\!\perp U$, as when P is experimentally controlled by the researcher. Similarly, instruments for (all) prices P could allow identification.

Of course, restricting the vector of demand shocks to affect demand for good j monotonically through a scalar $\mathcal{E}_j(U)$ involves a strong functional form restriction—one ruled out even by common parametric demand specifications like the multinomial probit, multinomial logit, or CES models.[7] And without this index restriction an experiment (or instrument for the prices P) generally will not suffice to allow identification of demand. As we will see in later sections of this chapter, identification of demand can still be obtained using additional sources of variation.

[7] The notable exception of demand that is linear in all demand shocks (own and other) points out a potentially unappreciated implication of a linear specification: a reduction in the dimension of exogenous variation required for identification.

2.5 Many Common Tools Fall Short

Our discussion has already suggested that a solution to the challenge of demand estimation can be obtained using instrumental variables, potentially with other sources of independent variation in prices or quantities. Most of our chapter focuses on specific parametric and nonparametric instrumental variables approaches to identification and estimation, using insights from several literatures in IO and econometrics. One may wonder whether the complexity of these strategies is necessary—specifically, whether simpler or more familiar tools of empirical economics might offer viable alternatives. In fact, many common empirical tools in economics—including various "research designs" useful for measuring (causal) effects of endogenous variables in other contexts—are not applicable to demand, at least without significant compromises or developments beyond the current methodological frontier.

We briefly discuss some of these alternative empirical approaches below. Because critical shortcomings are evident even in the idealized single-good supply and demand model given by (2.1) and (2.2), we focus primarily on this case to illustrate.

2.5.1 Controls, Including Fixed Effects

Because the problem of endogeneity arises from the presence of unobservables, a natural approach is to look for "controls" that will eliminate this problem. This can work, but the requirements are strong. First, the controls must absorb all effects of the unobservables on demand. Second, the controls cannot absorb all the variation in prices. This second requirement is a kind of exclusion restriction: there must be sufficient sources of price variation that are not included in the demand equation controls. Among other consequences is that the same set of controls cannot be used for both demand and supply estimation.

Note that if all demand shifters are observed and included in X in the demand equation (2.1), then we must observe a perfect fit/prediction of demand at all (X, P) (up to any measurement error in the quantity demanded). This is a strong requirement; but without it we are left with unobserved demand shifters and the problem of simultaneously determined (endogenous) prices.

Fixed effects are sometimes considered to be an attractive set of controls for precisely the reason that they might "control for everything." However, the requirements for a fixed effects strategy to solve the endogeneity problem are no different. It remains critical that while fixed effects control for everything affecting demand, they do not control for everything that affects price. This is a point worth emphasizing because the presence of latent demand shocks at the level of the product or market might suggest a fixed effects approach to controlling for these unobservables. However, a product-level fixed effect will not suffice if demand shocks vary by both product and market, as typically assumed.[8]

[8] Thus, for example, Nevo (2001) incorporates product fixed effects but still has demand shocks for each product×market combination, each representing the deviations of the product-level unobservable in that market from its overall mean.

Yet if the price of a given good is the same for all consumers in a given market (often this is the definition of "market"), a fixed effect for each product×market will leave no variation in price, making it impossible to measure demand elasticities or to connect demand to standard notions of aggregate welfare.

In a panel data set covering products within geographic markets across time, it is feasible to consider fixed effects for products across markets (held fixed over time) plus time effects (held fixed across product/markets). But for this to serve as a valid approach to the problem of endogenous prices, the resulting model must fit the data perfectly, up to measurement error in the observed quantities. And, again, the same set of fixed effects could not be used to estimate supply, because we require additional sources of supply-side price variation (not just measurement error).

2.5.2 Control Function

"Control function" approaches are popular and have close ties to IV approaches.[9] These approaches begin with "triangular" models of the form

$$T = F(Z, E_1) \tag{2.5}$$

$$Y = G(T, E_2), \tag{2.6}$$

where (2.6) represents an outcome equation of interest and (2.5) is a reduced form for the endogenous ("treatment") variable T appearing on the right-hand side of (2.6).[10] Here E_1 is assumed to be a scalar, with F strictly increasing in E_1. In such contexts, a control function can be used to treat the endogeneity of T in the outcome equation, allowing identification of G if E_2 is also a scalar (or, otherwise, identification of certain average effects).

One may be tempted to view demand as fitting this triangular structure, with price being the endogenous variable T affecting the quantity demanded, Y. But demand generally cannot be represented in this form. Even in the perfectly competitive single-good economy characterized by (2.1) and (2.2), the reduced form for the equilibrium price takes the form

$$P = R(X, W, U, V). \tag{2.7}$$

This does not take the form of (2.5). Critically, the right-hand side of (2.7) depends on all structural errors—here, the scalar demand shock U and scalar cost shock V. As demonstrated formally by Blundell and Matzkin (2014), only in very special cases (a linear model being one example) will the errors U and V enter the reduced form through a scalar index (not itself dependent on (X, W)), allowing for valid application of control

[9] Imbens and Newey (2009) (see also references therein) discuss nonparametric control function approaches. They note the typical failure of control functions in non-triangular systems, including in many classic simultaneous equations models.

[10] Throughout we use the term "reduced form" as it is defined in econometrics: a relationship in which an endogenous variables is expressed as a function of exogenous variables and structural errors.

function approaches. Without such functional form restrictions, however, the control function approach will not allow identification of demand.[11] The key problem is clear: in general a single control variate cannot eliminate the confounding effects of multiple unobservables, much less hold them fixed at particular values. Of course, this problem only becomes more severe when one acknowledges the presence of other related goods; in that case, each outcome (demand) equation depends on multiple endogenous prices, each of which generally depends on the demand shocks and supply shocks associated with all related goods.

2.5.3 Average Treatment Effects

The measurement of treatment effects is a major area of empirical economics that regularly deals with multiple unobservables and endogeneity, often focusing on estimation of average responses like a local average treatment effect ("LATE"). Such measures characterize responses averaged over certain values of the unobservables (Angrist and Imbens (1995)). A natural question is whether measurement of demand can be approached the same way (see, e.g., Angrist, Graddy, and Imbens (2000)).

In general, the answer is no. As we have suggested already, average responses of demand to price changes are of very limited economic interest. To illustrate this point as clearly possible, consider one of the first *ceteris paribus* counterfactuals taught to undergraduates: the effect of an excise tax, τ, on the equilibrium price $P^*(X, W, U, V, \tau)$ in the single-good supply and demand model of equations (2.1) and (2.2). Rewriting the marginal cost function in (2.2) in inverse form as "supply,"[12]

$$Q = S(W, P, V),$$

we teach (by graph and equation) that the change in the change in equilibrium price resulting from change in the excise tax depends on relative slopes of supply and demand, as measured by

$$\frac{\partial P^*(X, W, U, V, \tau)}{\partial \tau} = \frac{\frac{\partial S(W,P,V)}{\partial P}}{\frac{\partial S(W,P,V)}{\partial P} + \left|\frac{\partial D(X,P,U)}{\partial P}\right|}. \tag{2.8}$$

This ratio is not a LATE. By definition, a LATE averages over the latent variables; this is not the same thing as holding them fixed. Thus, a LATE approach cannot produce *ceteris paribus* counterfactual quantity of interest: the causal effect of the tax change.

But the limitations of a LATE in this example are typically much more severe than the distinction between the effect at a point and an average effect. Indeed, in some

[11] Petrin and Train (2010) and Kim and Petrin (forthcoming) describe special cases allowing use of a control function approach.

[12] For simplicity, here we assume upward sloping marginal cost and differentiability of the functions D and S. Note that elsewhere in this chapter we use the notation s to represent a vector of market shares. We trust our use of S for "supply" here will not cause confusion.

cases one might be interested in an average effect like a LATE for the treatment of a tax change. This would be equal to the left-hand side of (2.8) averaged over some distribution of (U, V). However, this cannot be determined from LATE estimates of demand (and supply). One could estimate separate local average derivatives of demand and supply with the LATE approach in this simple setting. However, a ratio of averages is not equal to the average ratio. Furthermore, because the weights for each local average depend on the instruments (see, e.g., Angrist and Imbens (1995) and Angrist, Graddy, and Imbens (2000)), the necessary use of different instruments to identify demand and supply averages will also imply different (and unknown) weights for each average. Thus, a LATE approach to demand (and supply) cannot identify even the average numerator and average denominator under a common measure.[13]

We emphasize that this example was chosen because it is perhaps the most elementary example of the kind of economic question that motivates demand estimation. It is not special in terms of the limitations of LATE. Almost any equilibrium counterfactual will involve interactions between demand and supply, leading to similar issues. Thus, it is not merely that a LATE approach to demand estimation fails to allow measurement of the quantity of primary economic interest; rather, it typically does not allow one to measure any well-defined average equilibrium counterfactual quantity.

Notably, the problems discussed in the preceding paragraphs are those arising in the simplest case—that of demand and supply for a single good with no complements or substitutes. With multiple goods, the shortcomings of a LATE approach also multiply. One then faces a system of equations characterizing the responses of demand to prices (recall section 2.3), even before turning to more complex counterfactuals. Multiple prices require multiple instruments, and the problems of LATE in handling different averages associated with different instruments become even more important. This only adds to the limitations of LATE for equilibrium counterfactuals. And in such cases (i.e., in the most common empirical setting in practice), it appears that LATE demand estimates could not produce even a local average own-price elasticity of demand.

Thus, although a LATE approach might embody the right set of compromises for many empirical settings, it is poorly suited to the economic questions that motivate demand estimation. Luckily, one can use different empirical tools depending on the empirical questions of interest. Much of our focus in what follows involves empirical approaches that allow one to model substantial unobserved heterogeneity (at the level of individuals, goods, and markets) while still permitting identification and estimation of the objects of economic interest in contexts involving demand.

[13] Of course, if the data offer adequate exogenous variation in the tax τ, one could estimate an average *ex post* effect without estimating demand—either from averages of equilibrium prices $P^*(X, W, U, V, \tau)$ or an instrumental variables LATE estimate. This serves as a reminder that demand estimation is typically motivated by a desire to answer questions—e.g., to infer oligopoly markups or provide policy advice on the implications of a proposed carbon tax—that require either an *ex ante* analysis or a counterfactual quantity that cannot be characterized by an average response of one scalar observable to an exogenous change in another.

2.6 Balancing Flexibility and Practicality

Although demand presents challenges that are absent in many empirical settings, all the "usual" challenges remain. One such challenge is finding empirical specifications that are both (a) sufficiently flexible to avoid strong *a priori* restrictions on the results and (b) sufficiently parsimonious to permit practical application. In some markets the number of closely related goods can be large—consider, for example, the set of all new automobile models, all computer models, all mutual funds, or all residential neighborhoods in a given city. Because the demand for a given good depends on the characteristics and prices of related goods, a demand system with J goods has J^2 price elasticities at each point. In many contexts, this will rule out even a linear specification of the demand equation (2.3).

Thus, even in cases where nonparametric estimation would be possible in principle, in practice it will often be necessary to impose restrictions in order to obtain an empirical model that is practical for the data available. Unsurprisingly, one can go too far in the pursuit of parsimony. Some of the simplest demand specifications (e.g., the CES, multinomial logit, multinomial probit) impose strong *a priori* restrictions on demand elasticities—and, therefore, on markups, pass-through, and other key quantities of interest—that are at odds with common sense and standard economic models.[14] Below we will discuss some of the strategies used in practice to strike a more attractive balance between parsimony and flexibility. One common strategy is to derive demand from a specification of consumer utility functions.

2.7 Demand or Utilities?

The most common approach to modeling demand estimation in the IO literature starts from a specification of consumer utilities. This is a matter of convenience rather than necessity. The primary goal in demand estimation is to obtain a quantitative representation of how quantities demanded respond to *ceteris paribus* variation in prices and other observables. This does not require a specification of utilities.[15] Indeed, a representation of consumer demand in terms of utility maximization exists only under certain restrictions on demand. And when one is willing to impose the conditions on demand that allow one to make valid welfare statements (see, e.g., Bhattacharya (2018) in the case of discrete choice demand), these may be assumed directly and utilized to construct welfare measures without actually specifying utility functions.

However, deriving demand from a specification of utilities can have significant practical advantages. A widely recognized (and often decisive) benefit is that such an approach can represent a demand system for many goods (and, thus, many own- and cross-price

[14] See, e.g., the discussion of in Berry, Levinsohn, and Pakes (1995) and, for the CES, in Adao, Costinot, and Donaldson (2017).

[15] In terms of consumer theory, one may view utilities as primitives and individual/aggregate demand as a derived object, or view individual choice rules (individual demand) as primitives, with utilities (and optimization) as a derived representation (see, e.g., Mas-Colell, Whinston, and Green (1995)).

elasticities) with a relatively small number of parameters (see, e.g., Berry (1994)). Typically, some of this parsimony comes from imposition of economically motivated restrictions. For example, researchers will often prefer to require that individual demand satisfy standard rationality conditions, which hold automatically when demand is derived from utility maximization. And even when focusing on market-level demand, an explicit connection to individual-level demand can often be usefully exploited in empirical work. At the consumer level, researchers often wish to impose certain economic restrictions or symmetry conditions that may be more easily formulated through utility functions.[16] Examples of such restrictions include:

- an assumption that heterogeneity in preferences over a given set of goods arises in part from consumer heterogeneity in tastes for the characteristics of the goods;
- an assumption that variation in the characteristics of good j alters the attractiveness of good j relative to others, but not the relative attractiveness of other pairs of goods;[17]
- an assumption that, all else equal, each consumer has a single marginal rate of substitution between any pair of product characteristics—say, price and quality—regardless of the name of the product.

Restrictions of these types are not without loss, but they can lead to specifications offering an attractive balance of parsimony and flexibility.[18]

3 Discrete Choice Demand

Building on the pioneering work of McFadden (1974, 1977), demand in many IO applications is formulated with a discrete choice model. Thus, although most key insights apply more broadly, much of our discussion will focus on discrete choice. In a discrete choice model each consumer selects exactly one of the options available to her. In most applications to demand, the options in the choice set are individual products. However, the class of discrete choice models is more general than it may seem. For example, one of the options could be to purchase both a Dodge Caravan and a Porsche 911, and another

[16] Compiani (2020) explores imposition of such restrictions without specifying utility functions.

[17] This assumption has similarities to Luce's independence of irrelevant alternatives ("IIA") axiom. However, the IIA assumption replaces "relative attractiveness" with "relative choice probabilities." Luce's IIA is a highly unnatural restriction (see, e.g., Debreu (1960), McFadden (1974), and Berry, Levinsohn, and Pakes (1995)), whereas the "relative attractiveness" assumption has intuitive appeal in the discrete choice context.

[18] Similarly, specifying a utility functions can facilitate the use of additional assumptions to answer questions like (a) whether larger values of an observed consumer characteristic make one good more attractive or (b) whether observed consumer characteristics like income or education alter preferences or serve as proxies for latent preference variation.

option (in the same demand system or, more likely, another) could be to purchase four boxes of cookies and a gallon of milk.[19] For simplicity, however, we will refer to each of the options as a good or product. Typically each consumer's choice set should include an option of the form "none of the above"—what we will call the "outside good." This is important. In a discrete choice model without an outside good, the market demand elasticity would always be zero; for example, when estimating demand for health insurance, a model with no outside good would imply that doubling all premiums would have no effect on the number of households with insurance. Note also that the choice probabilities implied by a discrete choice model can often also be interpreted as a demand system generated by continuous choices, as from a representative consumer with a taste for variety (see for example the review provided by Anderson, DePalma, and Thisse (1992)).

3.1 Random Utility Models

In most of the literature, discrete choice demand is represented with a "random utility" model. Let $j = 1, \ldots, J_i$ index the "inside goods" available to consumer i while $j = 0$ denotes the outside good. A consumer's choice set is characterized by J_i and a set χ_i, which may include observed characteristics of consumer i, observed characteristics (including prices) of the available goods, observed characteristics of the local market, and characteristics of the market or goods that are unobserved to the researcher. Each consumer i has a conditional indirect utility (henceforth, "utility") u_{ij} for good j. Consumer i knows her utilities for all goods and chooses the good yielding her the highest utility.

Consumer preferences are permitted to be heterogeneous, even when conditioning on any consumer characteristics included in χ_i. This heterogeneity is modeled by treating utilities as varying at random across consumers: given the choice set (J_i, χ_i), each consumer's utility vector $(u_{i0}, u_{i1}, \ldots, u_{iJ_i})$ is an independent draw from a joint distribution $F_u(\cdot \mid J_i, \chi_i)$.[20] Because a consumer's behavior depends only on her ordinal ranking of goods, below we will normalize the location and scale of each consumer's utility vector without loss of generality.[21] We assume that the distribution $F_u(\cdot \mid J_i, \chi_i)$ is such that

[19] See for example Gentzkow (2007), who notes that this approach can generate complementarities across the "primitive" products. There are practical limits to this flexibility, and one may want to impose cross-option restrictions when the distinct options in the choice set involve partially overlapping bundles.

[20] As a characterization of demand, the modeled randomness in utilities may be interpreted as reflecting heterogeneity in consumer's decision making, allowing for example mis-perception, inconsistencies, or non-optimizing behavior. See, e.g., the discussion and references in chapter 2 of Anderson, DePalma, and Thisse (1992). The choice of interpretation becomes important if one wishes to make welfare statements or counterfactual predictions associated with interventions that might alter consumers' decision rules.

[21] Normalizations of the location and scale of each consumer's utilities are without loss of generality with respect to behavior. However, different normalizations can have different implications for the interpretation of additional assumptions, including those used to justify certain welfare statements. See Bhattacharya (2018) for important results on standard aggregate welfare measures in random utility discrete choice models.

"ties" ($u_{ij} = u_{ik}$ for $j \neq k$) occur with probability zero.

We may then represent consumer i's choice with the vector $(q_{i1}, \ldots, q_{iJ_i})$, where

$$q_{ij} = 1\left\{u_{ij} \geq u_{ik} \ \forall k \in \{0, 1, \ldots, J_i\}\right\}.$$

Consumer-specific choice probabilities are then given by

$$\begin{aligned} s_{ij} &= E\left[q_{ij} \mid J_i, \chi_i\right] \\ &= \int_{\mathcal{A}_{ij}} dF_u\left(u_{i0}, u_{i1}, \ldots, u_{iJ_i} \mid J_i, \chi_i\right), \end{aligned}$$

where

$$\mathcal{A}_{ij} = \left\{(u_{i0}, u_{i1}, \ldots, u_{iJ_i}) \in \mathbb{R}^{J_i+1} : u_{ij} \geq u_{ik} \ \forall k\right\}.$$

To illustrate, consider an example with $J_i = 2$. Let p_j denote the price of good j and let

$$u_{ij} = \mu_{ij} - p_j$$

for $j > 0$, where (μ_{i1}, μ_{i2}) are drawn from a joint distribution $F_\mu(\cdot)$. Set $u_{i0} = 0$, normalizing the location of utilities. Figure 1 then illustrates the regions in (μ_{i1}, μ_{i2})-space leading consumer i to choose goods 0, 1, and 2. For example, only consumers for whom $\mu_{i2} - p_2 > 0$ prefer good 2 to the outside option. The dark grey region is the set of (μ_{i1}, μ_{i2}) combinations such that this holds and $\mu_{i2} - p_2 > \mu_{i1} - p_1$, i.e., the set $\mathcal{A}_{i2}$. Similarly, the light grey region corresponds to $\mathcal{A}_{i1}$. The choice probabilities for consumer i then correspond to the probability measure assigned to each region by $F_\mu(\cdot)$.

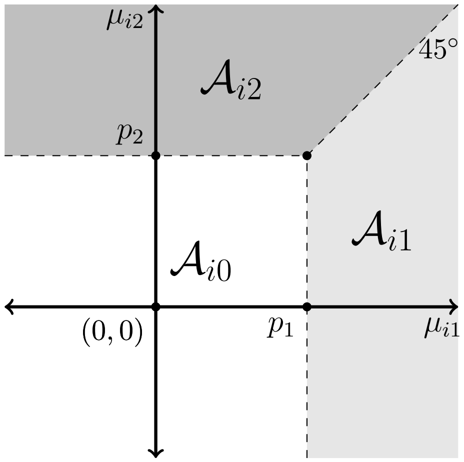


Figure 1: Choice regions for goods 0, 1, and 2.

3.2 The Canonical Model

Discrete choice demand models are frequently formulated using a parametric random utility specification such as[22]

$$u_{ijt} = x_{jt}\beta_{it} - \alpha_{it} p_{jt} + \xi_{jt} + \epsilon_{ijt} \tag{3.1}$$

for $j > 0$, with

$$u_{i0t} = \epsilon_{i0t}. \tag{3.2}$$

This formulation has many important components, which we discuss here in detail.

The notion of a "market" t is central to this formulation and will allow a precise characterization of the endogeneity problems inherent to demand estimation. In practice, markets are typically defined by natural combinations of time and geography—e.g., a given year in a given metropolitan area. Let $\mathcal{J}_t$ denote the set of products (inside goods) available to consumers in market t, and let $J_t = |\mathcal{J}_t|$. On the right-hand side of (3.1), p_{jt} represents the price of good j in market t, while $x_{jt} \in \mathbb{R}^K$ represents other observable characteristics of good j in this market.[23] The term ξ_{jt} is an unobserved factor—a demand shock—associated with good j and market t. The demand shock ξ_{jt} is often described as a measure of good j's unobserved characteristics. But this is more restrictive than necessary; ξ_{jt} can represent any combination of latent taste variation and latent product characteristics common to consumers in market t. For example, a high value of ξ_{jt} may simply indicate that consumers in market t have a high mean taste for good j. We let $x_t = (x_{1t}, \ldots, x_{J_t,t})$, $p_t = (p_{1t}, \ldots, p_{J_t,t})$, $\xi_t = (\xi_{1t}, \ldots, \xi_{J_t,t})$, and $\chi_t = (x_t, p_t, \xi_t)$.

Typically, one allows prices p_t to be correlated with ξ_t. One reason is that standard models of oligopoly competition imply that prices are endogenous; in particular, the equilibrium price of any good j in market t will depend on all components of x_t and ξ_t, as these alter the residual demand for good j. In addition, equilibrium prices are affected by latent shocks to marginal costs, which we typically expect to be correlated with demand unobservables. And when marginal costs are upward-sloping, this will imply dependence of equilibrium marginal costs (and, thus, prices) on demand shocks. Exogeneity of the remaining product characteristics x_t is often assumed, and we will do so in what follows. However, this is not essential. On the demand side, allowing endogeneity of additional characteristics is conceptually straightforward but will lead to

[22] See McFadden, Talvitie, and Associates (1977), Hausman and Wise (1978), McFadden (1981), Berry (1994), Berry, Levinsohn, and Pakes (1995), and a large literature that has followed. Observe that (3.2) provides the necessary location normalization of utilities. We emphasize that this normalization is without loss of generality with resect to the implied demand under the maintained assumption that the additive scalars ξ_{jt} fully capture the effects of unobservables at the product×market level. Such an assumption will sometimes be more plausible when controlling for systematic variation in the outside option as, for example, in Eizenberg (2014).

[23] We assume the first component of each x_{jt} is a one, absorbing the mean of ξ_{jt}. In some applications in which the same products appear in many markets, product dummies may be included. See, e.g., Nevo (2001).

more demanding instrumental variables requirements.[24]

The additive ϵ_{ijt} in (3.1) is most often specified as an i.i.d. draw from a standard type-1 extreme value distribution, yielding a mixed multinomial logit model. Alternatively, a normal distribution will yield a mixed multinomial probit.[25] The term "mixed" reflects the heterogeneity across consumers in the parameters α_{it} and β_{it} that characterize their marginal rates of substitution between the various observed and unobserved characteristics. Choice probabilities in the population reflect a mixture of the choice probabilities conditional on each possible combination of $(\alpha_{it}, \beta_{it})$. For example, in the case of mixed logit we have choice probabilities in the population (i.e., market shares) given by

$$s_{jt} = \int \frac{e^{x_{jt}\beta_{it} - \alpha_{it}p_{jt} + \xi_{jt}}}{\sum_{k=0}^{J_t} e^{x_{kt}\beta_{it} - \alpha_{it}p_{kt} + \xi_{kt}}} dF(\alpha_{it}, \beta_{it}; t), \tag{3.3}$$

where $F(\cdot;t)$ denotes the joint distribution of $(\alpha_{it}, \beta_{it})$ in market t. The latent taste parameters $(\alpha_{it}, \beta_{it})$ are often referred to as "random coefficients."[26]

To specify the joint distribution $F(\cdot;t)$, each component k of the random coefficient vector β_{it} is commonly specified as taking the form

$$\beta_{it}^{(k)} = \beta_0^{(k)} + \beta_\nu^{(k)}\nu_{it}^{(k)} + \sum_{\ell=1}^{L} \beta_d^{(\ell,k)} d_{i\ell t}. \tag{3.4}$$

Here $\beta_0^{(k)}$ is a parameter shifting all consumers' tastes for $x_{jt}^{(k)}$. Each $d_{i\ell t}$ represents a characteristic (e.g., demographic measure) of individual i, and each $\nu_{it}^{(k)}$ is a random variable with a pre-specified distribution (e.g., a standard normal). The parameters $\beta_d^{(\ell,k)}$ and $\beta_\nu^{(k)}$ govern the extent of variation in tastes for $x_{jt}^{(k)}$ across consumers with different demographic characteristics d_{it} or different taste shocks $\nu_{it}^{(k)}$. The distinction between $d_{i\ell t}$ and $\nu_{it}^{(k)}$ reflects the fact that each $d_{i\ell t}$ (or at least its distribution in the population) is assumed to be known. For example, in the case of demand for cars, one might specify that family size affects preference for large cars, in which case the actual distribution of family size in each market would allow the model to capture this source of latent preference heterogeneity in the population of consumers.[27] On the other hand, although we may also expect preference for fuel efficiency to vary in the population, there

[24] For example, given sufficient instruments, one may simply let p_t include all endogenous observables, as in Fan (2013). See also the discussion in Berry and Haile (2020).

[25] Independence of ϵ_{ijt} across goods j is not essential, and is often relaxed in the case of mixed probit. However, given the presence of the market-level demand shocks ξ_t and the random coefficients $(\alpha_{it}, \beta_{it})$, for the usual case in which prices are set at the market level (no individual-specific prices) it is standard to assume that each ϵ_{ijt} is independent of x_t, p_t, ξ_t.

[26] Early examples using random coefficients to generate random utility discrete choice models can be found in Quandt (1966, 1968). See also Quandt (1956).

[27] See, e.g., Goldberg (1995) and Petrin (2002).

may be no demographic measure whose distribution captures this heterogeneity. Such latent heterogeneity in preference for a product characteristic $x_{jt}^{(k)}$ can be captured by the random taste shocks $\nu_{it}^{(k)}$.

The treatment of the coefficient on price, α_{it} is similar. A typical specification of α_{it} takes the form

$$\ln(\alpha_{it}) = \alpha_0 + \alpha_y y_{it} + \alpha_\nu \nu_{it}^{(0)},$$

where y_{it} represents consumer-specific measures such as income that are posited to affect price sensitivity.[28] The variables included in y_{it} might overlap partially or entirely with d_{it}.

3.3 Why Random Coefficients?

In the canonical model, randomness in utilities reflects both the idiosyncratic "tastes for products" ϵ_{ij} and the random coefficients ("tastes for characteristics") $(\alpha_{it}, \beta_{it})$. One motivation for the latter can be illustrated by considering the same model without random coefficients:

$$u_{ijt} = x_{jt}\beta_0 - \alpha p_{jt} + \xi_{jt} + \epsilon_{ijt}. \tag{3.5}$$

Letting

$$\delta_{jt} = x_{jt}\beta_0 - \alpha p_{jt} + \xi_{jt}, \tag{3.6}$$

we can write (3.5) as

$$u_{ijt} = \delta_{jt} + \epsilon_{ijt}. \tag{3.7}$$

If the remaining stochastic terms ϵ_{ijt} are i.i.d. and independent of (x, p), products differ (up to realizations of ϵ_{ijt}) only in their "mean utilities" δ_{jt}.[29] This implies that choice probabilities depend only on these mean utilities. Likewise, price elasticities (own and cross) depend only on mean utilities. This is true not just for the multinomial logit model, but any additive random utility model of the form (3.7) with i.i.d. ϵ_{ijt}.

These are very restrictive implications. For example, they imply that any two goods with the same (or similar) market shares—no matter how they differ in other respects—will have the same (or similar) own-price elasticities, equilibrium markups, and cross-price elasticities with respect to any third good. These are not only strong restrictions, but properties that are contrary to economic models of differentiated products, where, for example, goods that are more similar tend to have larger cross-price elasticities.

To be clear, the problem is not just a lack of "realism," but the *a priori* restriction on key features like own and cross-price elasticities that motivate estimation of demand. Models of the form (3.7) impose very restrictive relationships between the *levels* of market shares and the matrix of own and cross-price *derivatives* and, therefore, on counterfactual

[28] Other functional forms are common in the literature. For example, it is common to specify price as entering in the form $\alpha \ln(y_{it} - p_{jt})$, following Berry, Levinsohn, and Pakes (1995).

[29] The term "mean utility" is standard but loose. Here the mean of u_{ijt} is equal to δ_{jt} plus the mean of ϵ_{ijt}, which need not be zero.

predictions. This is a bug, not a feature. These restrictions do not come from economics but from assumptions chosen for simplicity or analytical convenience. Models must, of course, abstract from reality, and finite samples require appropriate parsimony. But good modeling and approximation methods should aim to avoid strong *a priori* restrictions on the very quantities of interest unless those restrictions can be defended as natural economic assumptions.

Random coefficients are not the only way to avoid these restrictions. For example, the random terms $(\epsilon_{i1t}, \ldots, \epsilon_{iJ_t t})$ need not be specified as mutually independent. In the case with just a few products whose identity is constant across markets, a good alternative to random coefficients might specify an unrestricted covariance matrix for the ϵ_{it} vector. But in cases with more than a few products per market, or with products whose characteristics change across markets, random coefficients are attractive because they balance flexibility in key dimensions with tractability. Random coefficient specifications can be formulated using economics, building on the observations that real goods differ in multiple dimensions, and real consumers have heterogeneous preferences over these differences. Taking the case of automobiles, random coefficients on indicators for pickup trucks and for minivans enables the model to predict that different models of pickup trucks will be close substitutes, precisely because a consumer who likes one pickup truck will tend to be one with strong idiosyncratic taste for all pickup trucks. And even if the leading minivan and leading pickup truck have very similar market shares, the model can predict very different cross elasticities with respect to a third vehicle—say, an SUV targeted at families. Thus, as a matter of theory, random coefficients can introduce consumer heterogeneity along key dimensions of product differentiation. And a substantial empirical literature has demonstrated that in practice random coefficients can play a critical role in giving the demand specification sufficient flexibility to produce natural consumer substitution patterns.

As a practical matter, important questions include the measures included in x_t and the extent of heterogeneity modeled through random coefficients. In some cases, practical considerations may dictate selecting a set of observable characteristics viewed as most important or most subject to heterogeneity in preferences.[30] Depending on the data set, a specification with a very large number of random coefficients may yield imprecise estimates (particularly, of the parameters associated with the distribution of random coefficients),[31] or even numerical problems in estimation. A researcher then often faces the practical question of what product characteristics are modeled as having random coefficients. Should one choose just an index of "quality"? Just price? Dummy variables

[30] Gillen, Moon, and Shum (2014) and Gillen, Moon, Montero, and Shum (2019) propose a data-driven approach to selecting from a large set of observed characteristics assumed to affect only mean utilities.

[31] Of course, we often care more about the estimation error in our eventual counterfactual analysis than the statistical significance of the estimated random coefficient parameters *per se*. Furthermore, an imprecise estimate of the variance of a random coefficient, as may arise when instruments produce insufficient exogenous variation, should not be confused for evidence in favor of a degenerate coefficient.

indicating subsets (e.g., nests) of products?[32] Multiple observed characteristics (parts of x_t, p_t)? In practice, the choice must reflect the application and the available data. Economic considerations often suggest dimensions along which preference heterogeneity is likely to be most important for determining the consumer substitution patterns that drive own- and cross-price elasticities. But practical considerations such as sample size and available sources of exogenous variation may play a role as well. In many cases it may not be desirable to specify random coefficients on all components of (x_t, p_t). We will return to this issue below when discussing instrumental variables and identification.

4 Market-Level Data

In many applications the key data are observed at the market level. In such cases, one typically observes

- the number of goods J_t available to consumers in each market t;
- their prices and other observable characteristics p_t, x_t;
- their observed market shares, $\tilde{s}_{jt}$, typically measured as the total quantity of good j sold in market t divided by the number of consumers (e.g., households) in that market;
- the distribution of consumer characteristics (d_{it}, y_{it}) in each market; and
- possibly, additional variables w_t (e.g., cost shifters) that might serve as appropriate instruments

The standard approach to estimation of discrete choice demand from market-level data was developed in Berry, Levinsohn, and Pakes (1995), with many subsequent variations and extensions. Here we consider a slightly simplified version of their model with a non-random coefficient on price.[33] Thus, the random utility specification becomes

$$u_{ijt} = x_{jt}\beta_{it} - \alpha_0 p_{jt} + \xi_{jt} + \epsilon_{ijt} \tag{4.1}$$

for $j > 0$, with $u_{i0t} = \epsilon_{i0t}$. We follow Berry, Levinsohn, and Pakes (1995) in assuming that each ϵ_{ijt} is an i.i.d. draw from a standard type-1 extreme value (Gumbel) distribution,[34]

[32] This case covers the *nested logit* as a special case. See Ben-Akiva (1973), McFadden (1978) and, for the market-level IO context, Berry (1994).

[33] In practice, it is often important to allow for heterogeneity in price sensitivity. The variation of this model presented in section 6 illustrates a type of specification commonly used in practice, even in the case of market-level data. We present the more restrictive quasi-linear specification here to simplify exposition and make clearer the sources of key identification requirements.

[34] Setting the scale parameter of the Gumbel distribution to one normalizes the scale of utilities. Setting the location parameter to zero is also without loss due to the fact that adding the same constant to all utilities yields an equivalent representation of preferences.

and that each $\nu_{it}^{(k)}$ in (3.4) is an i.i.d. draw from a standard normal distribution.

Observe that (4.1) can be rewritten as

$$u_{ijt} = \delta_{jt} + \mu_{ijt} + \epsilon_{ijt}, \tag{4.2}$$

where we have defined

$$\delta_{jt} = x_{jt}\beta_0 - \alpha_0 p_{jt} + \xi_{jt} \tag{4.3}$$

and

$$\mu_{ijt} = \sum_{k=1}^{K} x_{jt}^{(k)} \left(\sum_{\ell=1}^{L} \beta_d^{(\ell,k)} d_{i\ell t} + \beta_\nu^{(k)} \nu_{it}^{(k)} \right). \tag{4.4}$$

Let $F_\mu(\cdot \mid x_t, \beta_d, \beta_\nu)$ denote the joint distribution of the stochastic terms $(\mu_{i1t}, \ldots, \mu_{iJ_t t})$ given $(x_t, \beta_d, \beta_\nu)$. Given our assumptions above, this distribution is known up to the parameters (β_d, β_ν).

Letting

$$\delta_t = (\delta_{1t}, \ldots, \delta_{J_t t}),$$

the market shares implied by the model take the form

$$\sigma_j(\delta_t, x_t, \beta_d, \beta_\nu, J_t) = \int \frac{e^{\delta_{jt}+\mu_{ijt}}}{\sum_{k=0}^{J_t} e^{\delta_{kt}+\mu_{ikt}}} dF_\mu(\mu_{it} \mid x_t, \beta_d, \beta_\nu, J_t) \tag{4.5}$$

for each good j. An important fact, demonstrated in Berry (1994), is that the demand system

$$\sigma(\delta_t, x_t, \beta_d, \beta_\nu, J_t) = (\sigma_1(\delta_t, x_t, \beta_d, \beta_\nu, J_t), \ldots, \sigma_{J_t}(\delta_t, x_t, \beta_d, \beta_\nu, J_t))$$

is invertible: given x_t, β_d, β_ν and any vector of nonzero market shares $s = (s_1, \ldots, s_{J_t})$ in market t such that $1 - \sum_{j>0} s_{jt} > 0$, there is a unique vector δ for market t such that

$$\sigma(\delta, x_t, \beta_d, \beta_\nu, J_t) = s.$$

4.1 The BLP Estimator

At the broadest level, an estimation strategy involves searching (or solving) for the parameters of the model that allow it to best fit the data. Let

$$\theta \equiv (\alpha_0, \beta_0, \beta_d, \beta_\nu)$$

represent all the parameters of the model. It will be useful to partition these as

$$\begin{aligned} \theta_1 &= (\alpha_0, \beta_0) \\ \theta_2 &= (\beta_d, \beta_\nu). \end{aligned}$$

In the literature, the elements of θ_1 are often referred to as the "linear parameters" and with θ_2 referred to as "nonlinear parameters."[35] Note that we can then rewrite the model's prediction of market shares (4.5) as

$$s_{jt} = \sigma_j(\delta_t, x_t, \theta_2, J_t).$$

Because identification of the model will rely on instrumental variables, it is natural to formulate an estimator using moment conditions. Berry, Levinsohn, and Pakes (1995) proposed a generalized method of moments (GMM) estimation approach that can be sketched as follows:

---

1. take a trial value of the parameters θ;
2. for each market t, "invert" the demand model at the observed market shares $\tilde{s}_t$ to find the unique vector $\xi_t \in \mathbb{R}^{J_t}$ such that, given the definition (4.3), $\sigma_j\left(\delta_t, x_t, \theta_2, J_t\right) = \tilde{s}_{jt}$ for all j;
3. evaluate the trial value θ using a GMM criterion function based on moment conditions of the form

   $$E[\xi_{jt}(\theta) z_{jt}] = 0,$$

   where $z_{jt} \supset x_{jt}$ is a vector of appropriate instrumental variables;
4. repeat from step 1 until a minimum is found.

More formally, let T denote the number of markets in the sample and let $N = \sum_{t=1}^{T} J_t$. The BLP estimator $\hat{\theta}$ is defined as the solution to a mathematical program:

$$\begin{aligned}
\min_{\theta} \quad & g(\xi\,(\theta))' \,\Omega\, g(\xi\,(\theta)) && (4.6)\\
\text{subject to} \quad & && \\
g(\xi(\theta)) &= \frac{1}{N} \sum_{\forall j,t} \xi_{jt}(\theta) z_{jt} && (4.7)\\
\xi_{jt}(\theta) &= \delta_{jt}(\theta_2) - x_{jt}\beta + \alpha p_{jt} && (4.8)\\
\log(\tilde{s}_{jt}) &= \log(\sigma_j(\delta_t, x_t, \theta_2, J_t)) && (4.9)\\
\sigma_j(\delta_t, x_t, \theta_2, J_t) &= \int \frac{\exp[\delta_{jt}(\theta_2) + x_{jt}\tilde{\beta}]}{1 + \sum_k \exp[\delta_{jt}(\theta_2) + x_{kt}\tilde{\beta}]} f_{\tilde{\beta}}(\tilde{\beta}|\theta_2) d\tilde{\beta}, && (4.10)
\end{aligned}$$

[35] Both sets of parameters alter demand nonlinearly, but the mean utilities δ_{jt} are linear in θ_1—a fact that can be exploited in computation of the estimator.

where Ω denotes the standard GMM weight matrix and $f_{\tilde{\beta}}(\cdot|\theta_2)$ denotes the joint density of $\tilde{\beta}_{it} \equiv \beta_{it} - \beta_0$, i.e., the consumer-specific components of the coefficients β_{it}. Computation and inference are discussed in section 4.4.

4.2 Instruments

Broadly speaking, estimation of demand requires observables that provide exogenous sources of independent variation in prices and quantities. In the case of market-level data, such variation must come from instrumental variables that are excluded from the relevant demand equations in an appropriate sense. The need to instrument for both prices and quantities may be counterintuitive: to estimate demand, we might think instruments for prices would suffice. As suggested in section 2, however, this is not the case.

The need for excluded instrumental variables beyond those for prices will be explained more formally in section 5. But this is easily seen in the BLP model by considering the hypothetical case of exogenous prices. Even in this case it is clear that the model cannot be estimated using only moments interacting the demand shocks ξ_{jt} with x_{jt} and p_{jt}: in a parametric model, identification requires at least as many moment conditions as parameters, and the parameters of the model include not only the coefficients $\theta_1 = (\alpha_0, \beta_0)$ on x_{jt} and p_{jt} in (4.3), but also the parameters θ_2 governing the variation in the random coefficients. Thus, additional moment conditions would be required.

Below we discuss several types of variables that can provide the necessary sources of exogenous variation in prices and quantities. That is, what types of observables can satisfy the requisite relevance and exclusion (conditional moment) conditions? An additional question is what functions of these observables most usefully play the roles of z_{jt} in the unconditional moment conditions whose sample analogs are given by (4.7). Thus, in this section we also discuss the approximation of "optimal instruments." One important lesson from the literature is that the use of (approximately) optimal instruments can greatly improve estimation precision.

4.2.1 Cost Shifters and their Proxies

A classic type of (excluded) instrument for estimating demand is an exogenous shifter of marginal cost, such as exogenous material costs, a tax, or tariff. In most models of supply, variation in marginal costs will be "passed through" to some extent. As long as these cost shifters are (mean) independent of latent demand shocks ξ_t, they can serve as appropriate instruments. Many natural cost shifters will vary across time and operate at the firm level; these will be most useful in applications where firms operate in few markets or when variation in "markets" has a substantial temporal component. Other measures such as location-specific distribution costs can provide variation even with global firms and no temporal variation. Similarly, if a firm producing for geographically distinct markets faces upward sloping marginal cost at the firm (product) level, the marginal cost associated with one market will be shifted by contemporaneous demand shifters in other markets.

Noisy measures of a producer's actual cost shifters can also serve as instruments. For example, the average wage level in a producer's labor market may not perfectly track the producers' labor costs but is nonetheless likely to be highly correlated with those costs. Thus, such wage measures can serve as instruments as long as they are uncorrelated with demand shocks conditional on the exogenous variables and consumer-specific measures (e.g., income and education) included in the demand model.

A less obvious type of proxy that can sometimes serve as an instrument for p_{jt} is the contemporaneous price of the same good in another geographic market (see, e.g., Hausman, Leonard, and Zona (1994), Hausman (1996), and Nevo (2001)). This is often referred to as a "Hausman instrument." The logic of this instrument is that even if we do not observe producer-specific cost shifters, variation in costs is likely present and at least partially responsible for variation in the prices a producer sets in all markets it serves. Thus, an observed price increase in market t' can signal a change in the producer's costs that also shifts its equilibrium price in market t. Although the logic of a Hausman instrument builds on that for a cost shifter, an important difference is that, outside a perfectly competitive model, prices reflect not just firm costs but also demand elasticities—something that depends on demand shocks. The excludability of Hausman instruments, therefore, requires close scrutiny. Taking the example above, the key assumption is that the price in market t' is (mean) independent of ξ_{jt} conditional on the exogenous x_{jt}. This would fail if the demand shocks ξ_{jt} and $\xi_{jt'}$ are correlated, for example through seasonal variation in demand that is not captured by the observable product characteristics. More generally, to use any proxy for an exogenous change in firm costs as an instrument, the proxy error must also be exogenous.

4.2.2 BLP Instruments

In addition to cost shifters and Hausman instruments, a third class of instruments involves the exogenous characteristics of competing products. In section 2 we explained that a fundamental challenge to estimation of demand is the fact that the demand for any one good j depends on the characteristics of all related goods. But while this elementary observation reveals a challenge, it can also provide a solution. In particular, as long as we can maintain the assumption of mean independence between x_t and ξ_t, the entire set x_{-jt} of competitors' product characteristics can serve as instruments (components of z_{jt}) creating exogenous variation in good j's market share. As just noted, these characteristics affect all quantities through the demand system. They also affect prices: the equilibrium markup for good j depends on the elasticity of the residual demand for good j, which again depends on the characteristics of all goods. Thus, exogenous characteristics of competing goods can provide exogenous variation in both prices and quantities.

These instruments—exogenous characteristics of competing goods—are often called "BLP instruments," following their use (along with other types of instruments) in Berry, Levinsohn, and Pakes (1995). In practice the BLP instruments are often strong shifters of both quantities and markups, particularly when used in good approximations to their optimal (efficiency maximizing) form (see section 4.2.5). We provide additional discussion of these instruments in section 5.4. Obviously, the validity of BLP instruments depends on their exogeneity (mean independence from the demand shocks). In some cases, it may be more natural to assume that at least some components of x_{jt} are chosen by firm j with knowledge of the demand shocks (or other shocks correlated with the demand shocks). In that case, it is clear that these product characteristics cannot be used as instruments, and an appropriate alternative strategy will be necessary.

4.2.3 Waldfogel-Fan Instruments

A fourth class of candidate instruments involves characteristics—e.g., average demographic measures—of nearby markets. In some applications these will act as exogenous shifters of equilibrium markups. Despite the reference to "other markets," the logic of these instruments is fundamentally different from that for Hausman instruments. In many applications, prices are set at a regional or "zone" level, with each zone covering more than one market.[36] In standard models, equilibrium prices for a zone will then depend on all factors affecting demand in the zone. For example, a market with a given distribution of income will be more likely (under zone pricing) to have high prices if it is adjacent to (in the same zone as) a high-income market than if it is surrounded by low-income markets. This will be true even if only some of the firms in the target market also operate in the nearby market (Fan (2013)). Thus, income in markets adjacent to market t will affect equilibrium markups in market t and can serve as an instrument for p_t, as long as the income measure is uncorrelated with the demand shocks ξ_t in market t. This will be most plausible when income in market t is already among the market observables x_t. Of course, other demographic measures may affect equilibrium pricing as well.

We refer to instruments relying on the distribution of consumer demographics as "Waldfogel instruments," since the logic follows the key insight emphasized by Waldfogel (2003): one's neighbors influence the types of products and prices one is offered.[37]

A related strategy can become available when firms compete in partially overlapping service areas. Following Fan (2013), who studied competition among newspapers, consider firms j and k whose service areas intersect in market t. If each firm sets a single price for its service area, then demographic characteristics throughout a firm's service area will affect its equilibrium pricing strategy. Consequently, demographics of all markets t' within firm k's service will affect firm j's markup. Thus, demographic characteristics anywhere within a competitor's service area can instrument for prices in

[36] See, e.g., Williams and Adams (2019) and DellaVigna and Gentzkow (2019).

[37] Of course income (or other market characteristics) may also shift the level of demand and therefore alter firm's marginal costs if those are upward sloping. Because the Waldfogel instruments affect markups even with constant marginal costs, one need not take a stand on whether marginal costs are upward sloping to support this IV strategy.

the market(s) served by firm j.[38]

4.2.4 Exogenous Measures of Market Structure

In some applications exogenous changes in market structure may occur over time. For example entry or exit of products will alter market shares directly and will alter equilibrium markups through the resulting changes in the intensity of competition (overall and locally in product space). When such entry and exit is exogenous, it can provide useful variation in both price and quantities. Measures of such variation are, in fact, just one form of the BLP instruments discussed above.

Another possible change in market structure is a change in firm ownership—a merger, spinoff, or possibly a change in the extent of common partial ownership. Profit maximization implies that such changes in ownership will alter the internalization of pricing externalities and, therefore, equilibrium markups.[39] When measures of such changes are independent of latent demand shocks (conditional on the other exogenous variables in the model), these can serve as instruments for prices.[40]

4.2.5 Optimal Instruments

Entirely separate from the question of which observables z_{jt} serve as exogenous instruments is the question of the optimal functions of these variables to use when transforming conditional moment restrictions like $E[\xi_{jt}|z_{jt}] = 0$ into unconditional moments like those defining the BLP estimator.[41] Intuitively, this is a question of what transformations of the exogenous variables yield the most useful variation for pinning down the parameters of the model. This can be particularly important when, as in many applications to differentiated products markets, there is a large number of excluded instruments which individually may have limited strength but which together can have strong effects on the relevant endogenous variables.

Formally, the question is what form of the instruments leads to asymptotic efficiency of the estimates. This is a standard problem in econometrics. For clarity, here we will write θ^0 to denote the true value of the parameter vector θ. Chamberlain (1986)

[38] With chains of partially overlapping service areas, demographics in the service areas of competitors to competitors could also serve as instruments. In practice, the power of such instruments would need to be checked.

[39] The extent to which variation in common partial ownership affects equilibrium pricing in practice is an area of active current research, relevant on its own but also to the potential use of common ownership measures as instruments.

[40] See, for example, Miller and Weinberg (2017).

[41] This question is also separate from the optimal linear weighting of a given set of moment conditions, which is already incorporated in the GMM objective function through the weighting matrix. When estimating demand alone, use of optimal instruments yields a just-identified model, making the GMM weight matrix irrelevant. This is not the case when estimating demand and supply jointly (see Conlon and Gortmaker (2020)).

considered the problem of optimal instruments under an assumed conditional moment restriction of the form $E[\xi_{jt}(\theta^0)|z_{jt}] = 0$. He showed, under certain assumptions, that the optimal instruments z^*_{jt} to use in an unconditional moment of the form $E[\xi_{jt}(\theta^0)z^*_{jt}] = 0$ are

$$z^*_{jt} = \Psi^{-1}_{jt} E\left[\frac{\partial \xi_{jt}(\theta^0)}{\partial \theta} \;\middle|\; z_{jt}\right], \tag{4.11}$$

where $\Psi_{jt} = E\left[\xi_{jt}(\theta^0)^2|z_{jt}\right].$

As in many nonlinear models, the optimal instruments are infeasible to compute, since they depend on both the unknown value θ^0 and the unknown distribution of the demand shocks that is implicit in the expectation operators. Several approaches to approximating the optimal instruments have been developed. While Berry, Levinsohn, and Pakes (1995) proposed an initial approach using low-order terms of a polynomial basis, improved options (improved basis functions or direct approximations of the expectations (4.11)) have been proposed by Berry, Levinsohn, and Pakes (1999), Reynaert and Verboven (2014), Gandhi and Houde (2020), and Conlon and Gortmaker (2020).

The consistent message from this literature is that use of (approximately) optimal instruments can substantially improve the precision of estimates. These alternative approximation approaches are discussed in detail by Conlon and Gortmaker (2020). Their associated PyBLP software incorporates many of these as options, along with the appropriate extensions for the case of joint estimation of demand and supply.

4.2.6 Evaluating Instruments

In the literature on linear IV models, it is nearly universal practice to report a "first-stage" regression of the endogenous variables on the instruments. This regression can produce various diagnostic statistics to evaluate whether the instruments are doing a "good job" of predicting the endogenous variables. Importantly, these diagnostics can provide tests for weak instruments. The literature on nonlinear IV does not always provide clear guidance on an appropriate analog to this important diagnostic exercise, but various authors have proposed some ideas to consider.

Some discrete choice models, like the logit and nested logit, produce expressions for δ_{jt} that are linear in all parameters, allowing for traditional first-stage regressions. For example, in the nested logit model we need instruments for price and the "within nest" share of a product (Berry (1994)), and we can run diagnostics on a traditional first stage. Salanie and Wolak (2019) derive a linear-in-parameters approximation to the BLP inverse share function, which could perhaps be used in a similar first-stage exercise. Some papers present other quasi-first stage exercises, for example regressions of prices and quantities on the exogenous variables. Motivated by the optimal instruments expression in (4.11), Berry, Carnall, and Spiller (1996) report an "ex-post" quasi first-stage by regressing $\partial \xi_{jt}(\hat{\theta})/\partial \theta$ on exogenous variables (with $\hat{\theta}$ denoting the estimated parameters.)

Further progress may be needed here, especially to consider weak instruments and related issues in the context of nonlinear models with multiple endogenous variables, multiple structural errors in each reduced form, and multiple instruments. Useful ideas may be found in Andrews and Guggenberger (2017), Andrews (2018), and Andrews and Mikusheva (2020). The separate topic of the local sensitivity of parameter estimates to IV assumptions is considered in Andrews, Gentzkow, and Shapiro (2017).

4.3 Using a Supply Side

In many cases, estimation of demand is motivated by questions defined by market counterfactuals involving both supply and demand. And even when one is focused exclusively on questions about demand, there can be substantial gains in precision from exploiting the additional restrictions that come from the equilibrium conditions of a supply model. Adding the supply side is relatively straightforward. Although almost any model of oligopoly supply can be considered, the most common is that implied by Nash equilibrium in a simultaneous price-setting game under complete information.

A key insight from BLP is that the system of multiproduct Nash price-setting first-order conditions can inverted to solve for the equilibrium level of marginal costs as function of (i) observed data and (ii) slopes of demand that are known once demand is identified. This insight extends easily to other static oligopoly models (e.g, quantity setting) and to nonparametric models.[42] The inverted first-order condition for each good j in these cases takes the form

$$mc_{jt} = \psi_j(s_t, p_t),$$

where the function ψ_j is known when the oligopoly game is specified and demand is identified. Thus, when demand is identified, identification of the equilibrium level of marginal costs (and therefore markups) follows immediately.

Since marginal costs can be recovered (up to sampling error) by inverting the equilibrium first-order conditions, standard arguments can allow identification and estimation of marginal cost functions as well. These functions can be of direct interest and they are essential for the identification of counterfactuals that alter the equilibrium quantities produced. But even if one's interest is limited to demand, one can add precision to the estimates of demand parameters by exploiting exogeneity assumptions involving shocks to marginal costs.

Suppose, for example, that we specify

$$mc_{jt} = c_{jt}\left(\mathrm{w}_{jt}, q_{jt}, \omega_{jt}, \gamma\right) = \mathrm{w}_{jt}\gamma_0 + \gamma_1 q_{jt} + \omega_{jt}, \tag{4.12}$$

where w_{jt} and ω_{jt} represent, respectively, observed and unobserved cost shifters associated with good j. Here we have introduced the new parameters $\gamma = (\gamma_0, \gamma_1)$ that govern the effects of cost shifters and quantity on marginal cost. Given any value of γ and the demand parameters θ, the inverted first-order conditions together with equation (4.12)

[42] Berry and Haile (2014) demonstrate nonparametric identification of marginal costs and cost functions for a large class of supply models. The approach here generalizes the use of imperfectly competitive first order conditions going back to Rosse (1970) and is inspired by the (somewhat different) use of multiproduct first-order conditions in Bresnahan (1987).

imply a unique value of ω_{jt}, so that we can write $\omega_{jt}(\sigma, \alpha, \beta, \gamma)$. Given any observables $\tilde{z}_{jt}$ that are assumed to be mean-independent of ω_{jt}, we now have additional moment conditions of the form

$$E\left[\omega_{jt}(\sigma, \alpha, \beta, \gamma)\, \tilde{z}_{jt}\right] = 0. \tag{4.13}$$

A typical assumption is that both w_t and x_t (the two may overlap) are mean independent of each ω_{jt}, yielding a large number of instruments that can be included in the supply-side instruments $\tilde{z}_{jt}$. Furthermore, (4.12) naturally models marginal cost as depending only on own-cost shifters and own-quantity. This implies that additional possible excluded instruments include the exogenous cost shifters and demand shifters for rival products. Thus, adding the supply side will often introduce few new parameters relative to the number of new moment conditions. Importantly, these supply moments depend not only on the cost parameters γ but also on the demand parameters. Thus, except in the case of just-identification, the supply moments (4.13) will provide information about demand parameters as well. In practice, this will often manifest through substantial improvements in the precision of the demand estimates—the parameters β_ν governing the heterogeneity in random coefficients, in particular—when one incorporates supply moments (see, e.g., Berry, Levinsohn, and Pakes (1995) and Conlon and Gortmaker (2020)).

When a model of supply leads to such overidentifying restrictions, it may be possible to use the satisfaction or failure of these restrictions to discriminate between alternative models of supply. This can be important in multiple ways. Hypotheses about firm "conduct" are often of direct interest. And to the extent that one is relying on the supply model for precise demand estimates, it would be valuable to have a way of evaluating the hypothesized model of firm behavior. This idea of discriminating between alternative models of firm conduct has its roots in the pioneering work of Bresnahan (1981, 1987). Although it is not possible to represent firm conduct as simply a parameter in a conjectural variations model, Berry and Haile (2014) have generalized key insights from that early literature to show that positing a model of firm conduct indeed provides falsifiable restrictions that can discriminate between alternative models of conduct, even without parametric specifications of demand or cost functions. The essence of their results is that there are many observable (or estimable) sources of variation in market conditions that alter equilibrium markups—differentially across different models of firm conduct. The comparative statics predictions of a given model of firm conduct typically will not align with the price variation observed in the data unless the hypothesized model is correct. We refer readers to Berry and Haile (2014) for a more formal discussion. Statistical testing procedures have recently been developed and applied by Backus, Conlon, and Sinkinson (2020) and Duarte, Magnolfi, Sølvsten, and Sullivan (2021).

4.4 Computing the BLP Estimator and Standard Errors

Berry, Levinsohn, and Pakes (1995) provided a computational algorithm for their estimator and for the associated standard errors. Their approach combined Monte Carlo approximation of the integrals defining market shares and demand elasticities with an algorithm similar to the sketch provided on page 23. This is an example of a nested-fixed-point algorithm, using a contraction mapping to solve a set of fixed-point equations for the demand shocks $\xi_t(\theta)$ that equate predicted and actual market shares in every market at each trial value of θ.

Whether estimating demand in isolation or jointly with supply, proper computation of the BLP estimator can be challenging. While many authors succeeded in implementing and customizing the BLP algorithm, naïve implementations can easily fail.[43] Over the last decade, several authors—notably Dubé, Fox, and Su (2012) and Conlon and Gortmaker (2020)—have aimed to modernize the approach to computing the BLP estimator by combining modern computing capabilities, new computational methods, and a set of "best practices" tailored specifically to computation of the BLP estimator. Important advances include computational power allowing improvement in procedures for approximation of the integrals defining market shares; Monte Carlo evidence yielding critical guidance on convergence tolerances; management of potential rounding, overflow, and underflow errors; new techniques for approximating optimal instruments; improved methods for computing pricing equilibria; and the use of modern solvers, often exploiting gradient-based optimization.

Conlon and Gortmaker (2020) discuss these and other important advances in detail and propose a modern version of the BLP nested-fixed-point algorithm.[44] Their paper also serves as an introduction to open-source software ("PyBLP") for implementing this approach, either for estimating demand in isolation or simultaneous estimation of demand and supply. Simulations in Conlon and Gortmaker (2020) illustrate relative advantages of alternative techniques available among the many options offered in the PyBLP software. We refer readers to Conlon and Gortmaker (2020) for details, extensive references, and advice on current best practices in computation of the BLP estimator, optimal instruments, standard errors, and equilibrium counterfactuals.

Regarding standard errors, note that the program on page 23 intentionally defines a set of moment conditions that permit the use of GMM inference techniques. There are four potential sources of variance in the moment conditions: (a) the data variance across products within markets, (b) the cross-market variance in data, (c) variance due to the finite sample of consumers used to construct the market shares $\tilde{s}_{jt}$ and (d) variance due to simulation draws (if any).

There is a small literature examining asymptotic issues that arise with the BLP es-

[43] Knittel and Metaxoglou (2014) document such possibilities.

[44] In an essential earlier contribution, Dubé, Fox, and Su (2012) examined potential problems with poorly formulated versions of the nested fixed point algorithm and proposed the alternative approach of applying standard specialized constrained optimization solvers to the program (4.6) defining the BLP estimator. Dubé, Fox, and Su (2012) showed that this program often can be reformulated to yield a form amenable to their "MPEC" (mathematical programming with equilibrium constraints) approach, particularly when first and second derivatives of the Lagrangian can be supplied. The authors have made code for this approach publicly available.

timator. Berry, Linton, and Pakes (2004b) discuss two related issues that arise with asymptotic approximations treating the number of products as growing large : (a) simulation error in the approximation of choice probabilities implied by the model; and (b) sampling error in the empirical market shares—sample means that are interpreted as approximations to the population means implied by the model. These issues are closely related, as some market shares must become small as the number of products per market grows. Berry, Linton, and Pakes (2004b) note that the nonlinear inversion for the mean utilities δ_t can cause the simulation and sampling errors to "blow up" as market shares become small. It is therefore important to control the simulation error as much as possible—by using a large number of simulation draws and/or importance sampling techniques—or else to avoid simulation entirely by using accurate numerical integration. Berry, Linton, and Pakes (2004b) discuss the need to account for simulation and sampling error when reporting estimation results, and provide formulas for doing so. To focus on the troublesome issue of small market shares, they present asymptotic variance results as the number of products, J, grows large.[45] Freyberger (2015) and Hong, Li, and Li (2020) provide general treatments of the asymptotics of the BLP-style estimators as the number of markets (and, if applicable, the number of simulation and sampling draws) grows large.

Aside from standard errors for parameter estimates, one will typically be interested in standard errors on counterfactual quantities of interest—e.g., a price change or welfare change under a hypothetical merger or counterfactual policy. Current practice is to construct such standard errors using either a parametric bootstrap or nonparametric bootstrap. In the former case, one simulates parameter draws from their normal asymptotic approximation and recomputes the implied quantity of interest for each draw (see, e.g., Nevo (2001)). In the latter case, one re-samples the data, re-runs the estimation procedure on each bootstrap sample, and computes the implied quantity of interest for each such bootstrap replication.

As a final issue, with sufficiently small consumer samples (relative to the number of products), one may observe market shares $\tilde{s}_{jt}$ for some goods that are equal to zero, even though the expected shares (the choice probabilities from the model) are strictly positive. This creates a problem for any estimation strategy relying on inversion of demand. Gandhi, Lu, and Shi (2019a) show that zero market shares nonetheless imply bounds on mean utilities, yielding an estimation approach that is valid in the presence

[45] Armstrong (2016) points out a number of pathologies that occur under particular assumptions as J grows large. For example, many models of competition will imply constant (or even zero) markups in the limit, hindering any IV strategy that relies on the induced variation in oligopoly markups. These results might be taken as a warning about how to develop useful asymptotic approximations when taking limits in the number of products; asymptotic results that assume perfect competition or constant markups will provide poor approximations for actual differentiated-products oligopoly markets in which markups are substantial and highly sensitive to the intensity of competition faced by each product. Note also that, although some types of instruments act only through their effects on markups, the essential role of BLP instruments from the perspective of identification (see section 5) is to provide exogenous variation in quantities conditional on prices.

of zero shares. Quan and Williams (2018) provide a more specialized solution to this problem when products appear in multiple markets.

5 Nonparametric Identification: Market-Level Data

We have presented a standard specification of demand derived from a random utility discrete choice model that incorporates a combination of parametric assumptions—on the functional form of utilities and on the distributions of latent heterogeneity across consumers. These specific choices are not essential: while some of the presented parametric choices lead to computational simplifications, the same approach generalizes easily to other parameterizations. For example, a researcher could relax the linearity of the mean utilities or replace one parametric distribution with another.

Because in practice all samples are finite, most empirical work relies to some degree on choices of functional form. This is true even of "nonparametric" estimation methods, which depend on the specification of, e.g., a kernel function or a finite-dimensional parametrization within a sieve sequence. There remains, however, the important question of whether parametric assumptions can be correctly viewed as finite-sample approximations or are in fact essential maintained assumptions.

In many simple empirical settings in economics, it is straightforward to show that parametric assumptions are not essential. That is, the quantities of interest are nonparametrically identified. For example, in the case of IV regression, it is not essential to assume the usual linear-in-parameters form. Rather, identification holds nonparametrically given appropriate instruments for any endogenous variables (see, e.g., Newey and Powell (2003)). Berry and Haile (2014) have demonstrated that the same is true for a flexible nonparametric model of demand for differentiated products, even when one has access only to market-level data.[46] Indeed, the essential requirements are precisely the instrumental variables conditions necessary for identification in the case of regression.

5.1 Insights from Parametric Models

Some useful intuition and insights about identification of demand can be gleaned by reviewing some familiar parametric models of discrete choice demand. In particular, these examples illustrate (i) the role of an index structure linking the unobservables ξ_{jt} to observables; (ii) the role that inversion of the demand system plays in yielding equations amenable to standard econometric tools; and (iii) the need for excluded instruments for all endogenous variables (prices and quantities) appearing in those equations.

[46] Here we focus on identification of demand. Berry and Haile (2014) also demonstrate nonparametric identification of marginal costs, cost shocks, and marginal cost functions, as well as discrimination between alternative models of supply.

5.1.1 Multinomial Logit

Consider a multinomial logit specification in which consumer i's utility from good j in market t takes the form

$$u_{ijt} = x_{jt}\beta - \alpha p_{jt} + \xi_{jt} + \epsilon_{ijt}.$$

A key feature of this model is the linear index

$$\delta_{jt} = x_{jt}\beta - \alpha p_{jt} + \xi_{jt}. \tag{5.1}$$

As is well known, each market share

$$s_{jt} = \frac{e^{\delta_{jt}}}{1 + \sum_k e^{\delta_{kt}}} \tag{5.2}$$

is a nonlinear function of the indices $\delta_{1t}, \ldots, \delta_{Jt}$. A key observation is that this map from indices to market shares is easily inverted; in particular,

$$\delta_{jt} = \ln(s_{jt}) - \ln(s_{0t}). \tag{5.3}$$

Substituting (5.3) into (5.1) yields an estimable equation for each good j of the form

$$\ln(s_{jt}) - \ln(s_{0t}) = x_{jt}\beta - \alpha p_{jt} + \xi_{jt}. \tag{5.4}$$

Estimation of this equation is straightforward (Berry (1994)). Indeed, although none of the demand (share) equations (5.2) is a regression equation, (5.4) is. Its identification requires one instrument for the one endogenous variable on the right-hand side. Thus, even a single market-level instrument for prices can suffice.

This basic "recipe" of "index-inversion-instruments" is transparent in the multinomial logit model. As we'll see below, it generalizes to other parametric and nonparametric models. To foreshadow elements of the fully nonparametric case, consider a more flexible semi-parametric model in which we (a) partition x_{jt} for each j as $\left(x_{jt}^{(1)}, x_{jt}^{(2)}\right)$, with $x_{jt}^{(1)} \in \mathbb{R}$, and (b) require that the multinomial logit structure hold only after conditional on the remaining exogenous observables $x_t^{(2)}$. This allows a much more flexible treatment of $x_t^{(2)}$ in the demand model. For example, the mean utility of good j could be affected by a fully nonparametric function of $x_{jt}^{(2)}$, or could even depend on x_{kj} for $k \neq j$. To consider identification in this more flexible model, we can simply fix $x_t^{(2)}$ at an arbitrary value and drop it from the notation, so that we have the random utility specification

$$u_{ijt} = x_{jt}^{(1)}\beta - \alpha p_{jt} + \xi_{jt} + \epsilon_{ijt} \qquad j = 1, \ldots, J.$$

The inverted demand equations, after dividing through by $\beta^{(1)}$, take the form

$$x_{jt}^{(1)} + \tilde{\xi}_{jt} = \frac{1}{\beta^{(1)}} \left(\ln(s_{jt}) - \ln(s_{0t})\right) + \frac{\alpha}{\beta^{(1)}} p_{jt}, \tag{5.5}$$

where $\tilde{\xi}_{jt} = \frac{\xi_{jt}}{\beta^{(1)}}$. Compared to (5.3), we still have a type of index—here $x_{jt}^{(1)} + \tilde{\xi}_{jt}$—on the left-hand side; and this index is still equal to a tightly parameterized function of markets shares and the price of good j. If we re-arrange one more time to write

$$x_{jt}^{(1)} = \frac{1}{\beta^{(1)}} \left(\ln(s_{jt}) - \ln(s_{0t})\right) + \frac{\alpha}{\beta^{(1)}} p_{jt} - \tilde{\xi}_{jt}, \tag{5.6}$$

we get something that now resembles a regression equation, with an additively separable error $\tilde{\xi}_{jt}$ on the right-hand side. This differs from a regression equation in one important way, however: the variable on the left-hand size is an exogenous product characteristic; i.e., it is mean independent of the "error term" $\tilde{\xi}_{jt}$. Writing the equation in this unusual way forms a connection to the more complicated models we discuss below. An implication of (5.6) is that, just as in the original fully linear multinomial logit model, only one excluded instrument z_{jt} is needed to identify this equation despite the presence of two right-hand-side endogenous variables, $(\ln(s_{jt}) - \ln(s_{0t}))$ and p_{jt}. In particular, with one such excluded instrument, we have the bivariate conditional moment restriction $E[\xi_{jt}|x_{jt}^{(1)}, z_{jt}] = 0$, which can identify the two parameters appearing on the right-hand side of (5.6).

5.1.2 Nested Logit

Consider now the more general nested logit model where the inverted demand system takes the form

$$\ln(s_{jt}) - \ln(s_{0t}) = x_{jt}\beta - \alpha p_{jt} + (1-\lambda)\ln(s_{j/g,t}) + \xi_{jt} \qquad j = 1, \ldots, J.$$

Here the subscript g denotes the nest (or "group") to which product j belongs. Compared to the multinomial logit, flexibility is added through the new coefficient λ on the within-group share $s_{j/g,t}$ of good j. Each equation of this inverted demand system again looks like a regression equation. However, we now need an instrument for the endogenous variable $\ln(s_{j/g,t})$ as well as for price. Note that $\ln(s_{j/g,t})$ is a particular function of the full share vector $(s_1, \ldots, s_J)$ implied by the parametric functional form.

Again, if we condition on $x_t^{(2)}$ and drop it from the notation, we can rewrite each equation of this inverted demand system as

$$x_{jt}^{(1)} + \tilde{\xi}_{jt} = \frac{1}{\beta^{(1)}} \big(\ln(s_{jt}) - \ln(s_{0t}) - (1-\lambda)\ln(s_{j/g,t})\big) + \frac{\alpha}{\beta^{(1)}} p_{jt} \tag{5.7}$$

On the left-hand side we have the same index derived under the multinomial logit; but on the right-hand side we have a more complicated function of market shares and price.

Again, the rearrangement has no effect on the set of instruments needed.

5.1.3 The BLP Model

Now consider the BLP mixed logit model discussed previously and suppose that there is at least one component $x_{jt}^{(1)}$ of the observed product characteristics x_{jt} that is specified as entering the model without a random coefficient. This is a restriction—and one we'll discuss further below; but in practice it is satisfied in almost all applications. As discussed earlier, the demand system has an inverse, usually written in terms of the mean utility vector δ_t or demand shocks ξ_t. However, as above we can also write the inverse for each good j in the form

$$x_{jt}^{(1)} + \tilde{\xi}_{jt} = \frac{1}{\beta^{(1)}} \tilde{\delta}_j \left(s_t, p_t, x_t^{(2)}, \theta \right). \tag{5.8}$$

On the left-hand side we have the same linear index in the previous examples. On the right-hand side, we haven an inverse market share function $\tilde{\delta}_j$, which we know to exist even though it lacks a closed form. This function depends on all of the model parameters and is a more complicated function of prices and market shares, all of which are correlated with the demand shock $\tilde{\xi}_{jt}$ in this equation. We explained earlier that with instruments z_{jt} comprising only the exogenous x_{jt} and an excluded instrument for p_{jt}, moment conditions of the form $E[\xi_{jt} z_{jt}] = 0$ will not suffice for identification. Equation (5.8) suggests why: we need instruments generating sufficient variation in the endogenous right-hand side variables s_t and p_t.

5.1.4 Index, Inversion, and Instruments

This review of how standard parametric models are identified brings out three recurring themes:

- demand shocks that enter through indices for each good;
- the presence of a one-to-one mapping between the indices and market shares, allowing inversion of the demand system;
- the application of instrumental variables to identify the components of the inverse demand.

We will see below that these same ideas allow one to demonstrate nonparametric identification of demand from market-level data. Following Berry and Haile (2014), we introduce a nonparametric index restriction to an otherwise very general demand model; the demand system is then inverted, yielding a set of inverse demand equations with one demand shock per equation. Standard IV conditions yield identification of these equations and, therefore, of the realized demand shocks. Identification of demand then becomes trivial, as the values of all variables in the demand system are known.

5.2 Nonparametric Demand Model

Without loss, we condition on a fixed number of inside goods J. Let demand for each good j in market t be given by

$$s_{jt} = \sigma_j \left(x_t, p_t, \xi_t\right) \qquad j = 1, \ldots, J. \tag{5.9}$$

Although we write s_{jt} on the left-hand side of (5.9), this measure of demand at the market level may be measured in quantities, market shares, or other one-to-one transformations of the demand vector (i.e., the quantities demanded).[47] As above, x_t represents all observed exogenous characteristics of the market and goods,[48] p_t represents the prices of all goods, and ξ_t represents the J-vector of demand shocks.

This representation of demand may be derived from a random utility discrete choice model in which the utilities $(u_{i1t}, \ldots, u_{iJt})$ are drawn from some unknown joint distribution $F_U(\cdot|x_t, p_t, \xi_t)$. However, the demand system (5.9) need not arise from a random utility specification, or even a discrete choice model. Indeed, thus far, the demand model is very general, with the only significant restriction being that the number of scalar demand shocks ξ_{jt} is J. To demonstrate identification, Berry and Haile (2014) require three main assumptions closely tied to our insights from the parametric examples.

5.2.1 A Nonparametric Index

Partition x_t as $(x_t^{(1)}, x_t^{(2)})$ where $x_t^{(1)} = (x_{1t}^{(1)}, \ldots, x_{Jt}^{(1)}) \in \mathbb{R}^J$. For each market t, define a vector of indices $\delta_t = (\delta_{1t}, \ldots, \delta_{Jt})$ where

$$\delta_{jt} = x_{jt}^{(1)} \beta_j + \xi_{jt}. \tag{5.10}$$

Assumption 5.1 (Index). *For all j, $\sigma_j \left(x_t, p_t, \xi_t\right) = \sigma_j \left(x_t^{(2)}, \delta_t, p_t\right)$.*

Assumption 5.1 requires that $x_{jt}^{(1)}$ and ξ_{jt} enter the nonparametric function σ only through the index δ_{jt}. This index requirement is an important nonparametric functional form restriction. As our examples above suggest, this type of restriction is implicit in parametric models widely used in the practice of demand estimation. Here, of course, the indices $(\delta_{1t}, \ldots, \delta_{Jt})$ are permitted to alter demand for each good j through a fully

[47] Berry, Gandhi, and Haile (2013) point out that transformations of quantities demanded to market shares, expenditure shares, or even purely artificial notions of "shares" can be useful for verifying conditions ensuring invertibility of the demand system (see section 5.2.2 below). However, if demand is invertible (or is identified) under one known injective transformation of quantities, it is under all such transformations as well.

[48] Any non-price market-level observables—e.g., average demographic measures—may be included in x_t.

nonparametric function σ_j. We will see below why this index structure is valuable,[49] and how it may be relaxed when one has access to "micro data" on individual choices rather than market-level data.

As in the parametric examples, the fact that demand shocks have no natural location or scale means that we may assume without loss of generality that $E[\xi_{jt}] = 0$ and $|\beta_j| = 1$ for all j. Thus, henceforth we work with indices of the form[50]

$$\delta_{jt} = x_{jt}^{(1)} + \xi_{jt}. \tag{5.11}$$

Furthermore, because the exogenous variables $x_t^{(2)}$ play no role in the identification argument, we will henceforth condition on an arbitrary value of $x_t^{(2)}$ without loss of generality and suppress $x_t^{(2)}$ in the notation.

5.2.2 Inverting Demand

In the parametric examples, a key step was inverting the demand system. We can use that strategy in the nonparametric model under a "connected substitutes" condition introduced by Berry, Gandhi, and Haile (2013).

Assumption 5.2 (Connected Substitutes)**.**
(i) $\sigma_k(\delta_t, p_t)$ is nonincreasing in δ_{jt} for all $j > 0$, $k \neq j$, and any $(\delta_t, p_t) \in \mathbb{R}^{2J}$;
(ii) for each $(\delta_t, p_t) \in$supp(δ_t, p_t) and any nonempty $\mathcal{K} \subseteq \{1, \ldots, J\}$, there exist $k \in \mathcal{K}$ and $\ell \notin \mathcal{K}$ such that $\sigma_\ell(\delta_t, p_t)$ is strictly decreasing in δ_{kt}.

Part (i) of Assumption 5.2 requires that goods be weak substitutes with respect to the indices: an improvement in the index δ_{jt} must weakly reduce the demand for other goods. This is automatic in a discrete choice setting in which δ_{jt} can be interpreted as an index altering good j's quality. While part (i) requires only weak substitution, part (ii) requires at least some strict substitution among goods $j = 0, 1, \ldots, J$—essentially, enough that there is no strict subset of goods that substitute only among themselves.

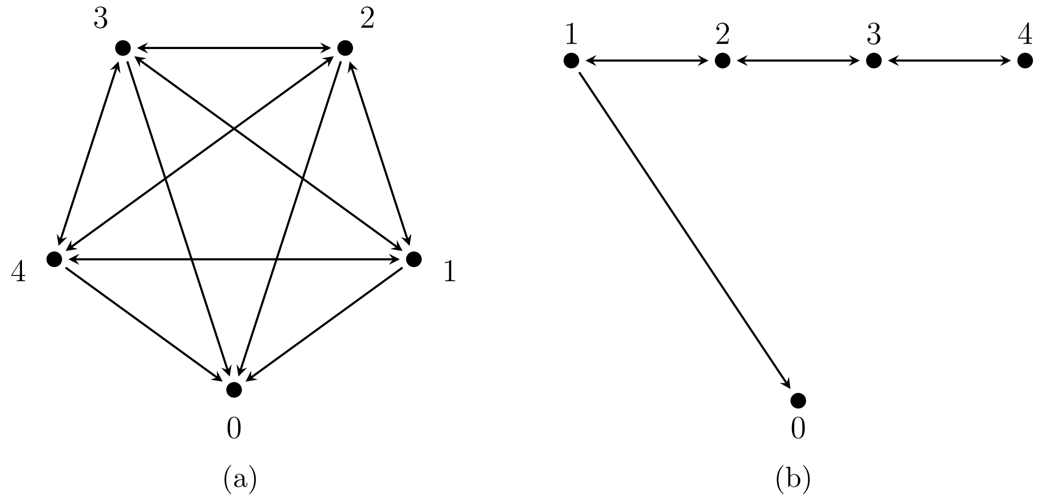
Berry, Gandhi, and Haile (2013) further show that part (ii) is equivalent to a certain notion of connectedness in the graph of the substitution matrix among goods. In particular, suppose we represent each good by a vertex and construct a directed edge from good j to good k if j strictly substitutes to k—i.e., if a reduction in δ_{jt} would lead to an increase in the demand for good k. Part (ii) requires (with appropriate quantifiers) that this directed graph exhibit a path from each node j to the node 0 associated with the outside good. Figure 2 illustrates this graph for 2 classes of standard discrete choice models in which δ_{jt} shifts the utility for good j without affecting the utility from other goods; panel (a) shows the graph for standard random utility models with horizontal

[49] Berry and Haile (2014) show that the additive separability of (5.1) is not essential, although relaxing separability requires strengthening the "relevance" condition on instrumental variables, just as in the case of regression models (Chernozhukov and Hansen (2005)).

[50] If the sign of β_j is not known *a priori*, it is easily determined under our assumptions.

differentiation (e.g., multinomial logit or probit, nested logit, mixed logit); panel (b) shows that for models of vertical differentiation (e.g., Bresnahan (1981)). In both cases we see that from the vertex associated with any good $j > 0$, there is a directed path to the vertex associated with the outside good, as required by the connected substitutes conditions.

Figure 2: Substitution in Standard Discrete Choice Models



Directed graphs of the substitution matrix for standard discrete choice models, with $J = 4$ inside goods. Panel (a): standard random utility models of horizontal differentiation, such as the multinomial logit, multinomial probit, nested logit, mixed logit/probit. Panel (b): the pure vertical model with an outside good. From each vertex associated with an inside good there is a directed path to the vertex associated with the outside good.

Berry, Gandhi, and Haile (2013) demonstrate satisfaction of this condition in a wide range of demand models and demonstrate that, because invertibility of demand is ensured whenever the connected substitutes conditions hold for some injective transformation of the demand system, invertibility can be demonstrated by these conditions even in some cases in which goods are complements. We refer readers to Berry, Gandhi, and Haile (2013) for additional examples and discussion. For our purposes, the key implication is that for all demand vectors s_t such that $s_{jt} > 0$ for all j, there exists an inverse demand system taking the form

$$\delta_{jt} = \sigma_j^{-1}(s_t; p_t) \qquad j = 1, \ldots, J. \tag{5.12}$$

5.3 Identification via Instruments

We will see that with equation (5.12), identification of demand will follow from availability of instruments satisfying the same conditions required of instruments for identification of regression models. In the case of nonparametric regression we are interested in an equation of the form[51]

$$y = \Gamma\left(x\right) + \epsilon, \tag{5.13}$$

where $x \in \mathbb{R}^K$. Newey and Powell (2003) showed that given instruments z satisfying the mean independence condition $E[\epsilon|z] = 0$, a necessary and sufficient condition for identification of the regression function Γ is a standard "completeness" condition: that in the class of functions $B(\cdot)$ on $\mathbb{R}^K$ such that $E[B(x)|z]$ is finite, the only function B such that $E[B(x)|z] = 0$ almost surely is a function that maps to zero almost surely on its domain. This "completeness" condition is, thus, the nonparametric analog of the standard rank condition for linear regression. It is the formal "relevance" requirement on the instruments, defining precisely what it means for them to provide sufficient exogenous variation in the regressors x.

To connect this to demand, observe that we may re-arrange each equation of (5.12) as

$$x_{jt}^{(1)} = \sigma_j^{-1}(s_t; p_t) - \xi_{jt}, \tag{5.14}$$

yielding a form similar to (5.13). Unlike the regression equation, here we have an exogenous variable $x_{jt}^{(1)}$—indeed, a variable that is an essential instrument—on the left and all endogenous variables, (s_t, p_t), on the right. Nonetheless, a very similar argument demonstrates that with excluded instruments z_t, identification of the inverse demand function σ_j^{-1} holds on the same pair of conditions ensuring identification in the case of regression.

Assumption 5.3 (Instruments)**.** *(i) For all $j = 1, \ldots, J$, $E[\xi_{jt}|z_t, x_t^{(1)}] = 0$ almost surely. (ii) For all functions $B\left(s_t, p_t\right)$ with finite expectation, if $E\left[B\left(s_t, p_t\right)|z_t, x_t^{(1)}\right] = 0$ almost surely then $B\left(s_t, p_t\right) = 0$ almost surely.*

Lemma 1. *Under Assumptions 5.1–5.3, for all $j = 1, \ldots, J, \sigma_j^{-1}$ is identified on the support of (s_t, p_t).*

We refer readers to Berry and Haile (2014) for the proof, which follows that of Newey and Powell (2003) closely. Observe that when each function σ_j^{-1} is known, each ξ_{jt} can be inferred immediately from prices, market shares, and (5.14). With each ξ_{jt} known, identification of demand follows directly from equation (5.9).

Theorem 5.1 (Berry and Haile (2014))**.** *Suppose (s_t, x_t, p_t, z_t) are observable and that Assumptions 5.1–5.3 hold. Then for all j, the demand function σ_j is identified.*

[51] We abuse notation here to follow standard convention, letting y, x, and ϵ denote the components of a regression model (with z denoting an excluded instrument). However, none of these components should be confused with similarly denoted elements of the demand models we discuss.

5.4 Discussion

A broad interpretation of this identification result is that, despite the substantial challenges discussed in section 2, the main requirement for identification of demand is a completely standard one: availability of suitable instruments for the endogenous variables. On one hand, this is comforting. On the other hand, the result raises several natural questions about the precise instrumental variables requirements and whether further structure on the demand model might relax these requirements.

5.4.1 Why $2J$ Instruments?

Theorem 5.1 required instrumental variables conditions on both the exogenous product characteristics $x_t^{(1)}$ and excluded instruments z_t. The product characteristics $x_t^{(1)}$ have dimension J by construction. And a necessary condition for satisfaction of the completeness condition is that z_t have dimension of at least J. Candidates for z_t include those discussed in section 4.2: cost shifters, proxies for cost shifters (e.g., "Hausman instruments"), and markup shifters such as "Waldfogel instruments" or exogenous measures of competition. The need for these instruments is quite intuitive: to learn about demand, we need exogenous variation in the J-dimensional price vector p_t. Less intuitive is the need for the J-dimensional exogenous product characteristics $x_t^{(1)}$ as well.

This requirement ties directly to our discussion in section 2 of the fact that a J-dimensional unobserved vector ξ_t appears as a determinant of the demand for each good. Inversion of the demand system is a "trick" that yields a set of equations

$$x_{jt}^{(1)} = \sigma_j^{-1}(s_t; p_t) - \xi_{jt}, \tag{5.15}$$

each with one demand shock, as needed for application of standard instrumental variables arguments. However, each of these inverted demand equations features endogenous variables $(s_t, p_t) \in \mathbb{R}^{2J}$ on the right-hand side. More intuitively, observing that the function σ_j^{-1} above must be strictly increasing in s_{jt}, we can rewrite (5.15) as

$$s_{jt} = h_j\left(s_{-jt}, p_t, x_{jt}^{(1)}, \xi_{jt}\right)$$

for some function h_j. This formulation suggests that $x_{jt}^{(1)}$ can act as an "instrument for itself," but that we still need J instruments for prices and $J-1$ instruments for the endogenous quantities s_{jt}.

Putting this differently, we know that to quantify demand, we must pin down how the quantity demanded of each good j changes when one price varies and all others are held fixed. This indicates a need for exogenous shifters of prices. But we know (recall section 2.1) that exogenous variation in prices is not sufficient. We must also hold the demand indices δ_t fixed. This is not obviously feasible because the demand shocks entering δ_t are not observed. However, by exploiting the bijection between the vector δ_t and the vector of market shares s_t (all else fixed), we can indirectly control the value of δ_t by controlling the share vector s_t. This requires additional instruments for the J-vector of market shares.

The need to instrument for market shares can be seen in our parametric examples as well. There we saw that as we increased the flexibility of the demand model, the inverse demand equation used for estimation involved not only prices but also the market share vector: market shares entered with no unknown parameters in the case of the multinomial logit, but as we added flexibility to the parametric model, additional moment conditions were required to estimate the inverse demand functions. In the limiting case of a fully nonparametric model, we require instruments for all prices and quantities (market shares).

5.4.2 Why BLP Instruments?

The elements of the J-vector $x_t^{(1)}$ are examples of the so-called BLP instruments. Collectively, they have dimension J by construction. They shift prices through the supply equilibrium. But at any given price vector, the observables $x_t^{(1)}$ also directly affect demand. Thus, these may serve the role of the required instruments for market shares. In fact, these instruments are essential to identification of the nonparametric demand model in the case of market-level data. To see this, let us suppose that prices in each market were set exogenously at levels independent of the demand shocks ξ_t. We still have the fundamental problem that demand for each good depends on the J-vector ξ_t. Inverse demand still takes the form

$$\xi_{jt} = \sigma_j^{-1}\left(s_t; p_t\right) - x_{jt}^{(1)}$$

or, equivalently,

$$s_{jt} = h_j\left(s_{-jt}, p_t, x_{jt}^{(1)}, \xi_{jt}\right).$$

Identification of h_j or σ_j^{-1} will, intuitively, still require $J-1$ instruments for s_{-jt}. Cost shifters and markup shifters are of no help in this case: because they affect market shares only through prices, they provide no variation in demand beyond that already accounted for directly through p_t. And we have already conditioned on (held fixed) $x_t^{(2)}$, leaving no variation to exploit. Thus, the demand shifters $x_{-jt}^{(1)}$ are the only possibility in the "menu" of candidate instruments discussed above.[52]

More simply, we have seen that in the most general model we need independent variation in all J shares and all J prices. So no matter how much variation we can generate in prices, we need something else that moves all shares at any given price vector. The BLP instruments are the only candidates in our setting of market-level data.

[52] Additional types of observables or additional assumptions could offer alternative strategies. See, for example, sections 5.4.4, 6, and 7.

5.4.3 Why the Index?

Although the preceding discussion explains why the variation provided by BLP instruments is critical, it may be less clear why these instruments are valid, even given the exogeneity required by Assumption 5.3. The power of these instruments to shift market shares is clear. But how is it that observables entering the demand system can end up satisfying the relevant exclusion restriction? In what sense are these observables properly excluded? The answer to this question emphasizes one role that the index restriction is playing in the identification argument.

Recall that we conditioned on $x_t^{(2)}$, so the inverse demand function associated with the index δ_{jt} is really

$$\sigma_j^{-1}\left(s_t, p_t; x_t^{(2)}\right)$$

In terms of the identification argument, conditioning on $x_t^{(2)}$ really means that $x_t^{(2)}$ "instruments for itself." The index restriction is what leaves $x_t^{(1)}$ out of the function σ_j^{-1}, making $x_t^{(1)}$ available as instruments for shares—in particular, allowing $x_{-jt}^{(1)}$ to serve as instruments for s_{-jt}.

This feature of the index is closely connected to a restriction that is often so natural that it is assumed without comment: that characteristics of a good alter its utility but not the utility from other goods. In the canonical random utility model, this applies to all components of x_{jt}, p_{jt}, ξ_{jt}, leaving x_{-jt} out of the inverted demand equation expressing ξ_{jt} as a function of market shares, prices, and the exogenous observables x_{jt}. Although we do not require a specification of utilities, the index structure leaves the role of $x_t^{(2)}$ entirely free while preserving this same natural feature with respect to the components of $x_t^{(1)}$.

5.4.4 Further Restrictions and Tradeoffs

The prior discussion makes clear the essential roles of the index restriction, BLP instruments, and additional instruments for prices—at least in the absence of further restrictions on the model or additional observables. Here we give just two examples illustrating how additional structure can soften the IV requirements.

Suppose that, instead of $x_{jt}^{(1)}$, it is the price of good j that enters only through good j's index. Then

$$\delta_{jt} = \xi_{jt} - \alpha p_{jt},$$

where we now set $\alpha = 1$ without loss of generality. Treating x_t fully flexibly and conditioning on it, the inverse demand equations would then take the form

$$\xi_{jt} = \sigma_j^{-1}\left(s_t\right) + p_{jt}$$

or

$$p_{jt} = -\sigma_j^{-1}\left(s_t\right) + \xi_{jt}.$$

In this case we would require only J instruments, providing exogenous variation in s_t that need not be independent of prices. In this case, cost shifters, Hausman instruments, Waldfogel instruments, etc. are all candidates.

As a second example, suppose that both $x_t^{(1)}$ and prices enter only through the indices, as

$$\delta_{jt} = \xi_{jt} + x_{jt}^{(1)}\beta - \alpha p_{jt},$$

where we may again set $\alpha = 1$ without loss. The inverse demand equations then have the form (suppressing $x_t^{(2)}$)

$$\xi_{jt} = \sigma_j^{-1}\left(s_t\right) - x_{jt}\beta + p_{jt}$$

or

$$p_{jt} = -\sigma_j^{-1}\left(s_t\right) + x_{jt}\beta + \xi_{jt}$$

Again, instruments are required only for the market share vector s_t, and the BLP instruments x_{-jt} are now available options in addition to those discussed above.

These are just two examples illustrating trade-offs between functional form restrictions and IV requirements. This points out one role that even nonparametric functional form restrictions can play in practice: filling in the gap between the exogenous variation that may be available in a given data set and that which would be needed to discriminate between all nonparametric demand systems. Of course, some types of restrictions—e.g., adding symmetry, exchangeability, nesting, etc.—can arise as natural economic restrictions. And in some cases, better data will provide an avenue for relaxing the IV requirements. We turn next to a leading example of this: consumer-level data.

6 Micro Data, Panels, and Ranked Choices

6.1 Micro Data

In the context of demand estimation, the term "micro data" typically refers to a setting where a researcher observes individual consumer characteristics d_{it} matched to the choices q_{it} (e.g., vectors of quantities purchased) of each consumer. With market-level data, the observables can reveal only the marginal distributions $F_d\left(d_{it}\right), F_q\left(q_{it}\right)$ of consumer-specific observables (e.g., demographics) d_{it} and consumer-level choices q_{it} (conditional on market-level observables) in each market. Micro data can reveal their joint distribution $F_{dq}\left(d_{it}, q_{it}\right)$. This clearly provides useful information about how the observables d_{it} alter individual choices. For example, this could allow assignment of each consumer type—each value of d_{it}—to its own market, making the model completely flexible with respect to the effects of d_{it}. However, if one retains a more standard notion of market (e.g., based on a combination of time and geography), micro data provide a panel structure: observed outcomes for many individual consumers within each market. A key benefit—even if one is ultimately interested only in the market-level demand faced by firms—is that one can then exploit variation across consumers within each market, where the product×market-level demand shocks are fixed.

McFadden's classic work on demand for transportation modes (McFadden, Talvitie, and Associates (1977)) is an example of estimation from micro data. Other prominent examples involve demand for hospitals (e.g., Capps, Dranove, and Satterthwaite (2003), Ho (2009)), retail outlets (e.g., Burda, Harding, and Hausman (2015)), residential locations (e.g., Bayer, Ferreira, and McMillan (2007), Diamond (2016)), automobiles (e.g., Goldberg (1995), Petrin (2002)) and schools (e.g., Neilson (2020)). In these cases, observed demographic measures and geographic locations often play important roles. Other examples of consumer-level observables include product-specific advertising exposure (Ackerberg (2003)), consumer-newspaper ideological match (Gentzkow and Shapiro (2010)), or the match between household demographics and those of a school or neighborhood (Bayer, Ferreira, and McMillan (2007), Hom (2018)).

As we will see below, one significant advantage of micro data is the potential for within-market variation to lessen (but not eliminate) reliance on instrumental variables. Indeed, micro data can both reduce the number of instrumental variables necessary for identification and make new kinds of instrumental variables available.

To consider parametric models, we again focus on a mixed logit specification, with conditional indirect utilities of the form

$$u_{ijt} = x_{jt}\beta_{it} - \alpha_0 p_{jt} + \xi_{jt} + \epsilon_{ijt},$$

where

$$\beta_{it}^{(k)} = \beta_0^{(k)} + \sum_{\ell=1}^{L} \beta_d^{(\ell,k)} d_{i\ell t} + \beta_\nu^{(k)} \nu_{it}^{(k)}.$$

We can rewrite this as

$$u_{ijt} = \delta_{jt} + \mu_{ijt}(\nu_{it}; \beta_d, \beta_\nu) + \epsilon_{ijt}, \tag{6.1}$$

with

$$\mu_{ijt}(\nu_{it}; \beta_d, \beta_\nu) = \sum_{k=1}^{K} x_{jt}^{(k)} \left(\beta_\nu^{(k)} \nu_{it}^{(k)} + \sum_{\ell=1}^{L} \beta_d^{(\ell,k)} d_{i\ell t} \right) \tag{6.2}$$

and

$$\delta_{jt} = x_{jt}\beta_0 - \alpha_0 p_{jt} + \xi_{jt}. \tag{6.3}$$

McFadden, Talvitie, and Associates (1977) referred to δ_{jt} as the "alternative-specific" constant, which was held fixed in certain policy counterfactuals. The modern IO literature emphasizes the dependence of these "constants" on prices and other observed and unobserved factors. Of course, demand elasticities (and other key aspects of demand) are defined by responses to variation in one determinant of δ_{jt} while all other determinants are held fixed.[53]

As discussed in section 2, the latent demand shocks in these constants introduce the challenges of simultaneity/endogeneity. There was at one point perhaps some confusion as to whether the simultaneity problem arises in the micro-data context, reflecting the observation that the individual consumer does not "cause" the price. But this observation has no bearing on the simultaneity problem that arises from unobservables at the level of the product or market. We will see below that micro data can provide partial solutions to the simultaneity challenges. However, as our discussion below makes clear, the problems of simultaneity/endogeneity stem not from the level of aggregation of the data but from the presence of market×product-level demand shocks whose effects are confounded with those of market×product-level prices. Addressing this endogeneity problem generally will still require cross-market variation and instruments for prices.

With the specification above, choice probabilities for each consumer i take the form

$$s_{ijt} = \int \frac{\exp\{\delta_{jt} + \mu_{ijt}(\nu_{it}; \beta_d, \beta_\nu)\}}{\sum_{k=0}^{J_t} \exp\{\delta_{kt} + \mu_{ikt}(\nu_{it}; \beta_d, \beta_\nu)\}} dF_\nu(\nu_{it}) \tag{6.4}$$

for each good $j = 0, 1, \ldots, J_t$. Let $j(i)$ denote the good selected by consumer i in market t. Substituting $j(i)$ for j in (6.4) gives the likelihood contribution of consumer i's choice as a function of parameters $(\delta, \alpha_y, \beta_d, \beta_\nu)$, where we have now defined $\delta = \{\delta_t\}_{t=1}^T$ and treated it as a parameter vector. Although this likelihood would need to be approximated by simulating from the distribution F_ν, this immediately suggests an estimation approach for these parameters, at least when the number of observed consumers per good is large in each market.[54] In particular, one could estimate $(\delta, \beta_d, \beta_\nu)$ by maximizing the product of these (simulated) likelihoods over all consumers,

$$\mathcal{L}(\delta, \beta_d, \beta_\nu) = \prod_{i,t} \int \frac{\exp\{\delta_{j(i)t} + \mu_{ij(i)t}(\nu_{it}; \beta_d, \beta_\nu)\}}{\sum_{k=0}^{J_t} \exp\{\delta_{kt} + \mu_{ikt}(\nu_{it}; \beta_d, \beta_\nu)\}} dF_\nu(\nu_{it}). \tag{6.5}$$

Of course, to answer most economic questions one will also need to estimate the parameters α_0 and β_0 in (6.3). The simplest approach is to run a second-step linear IV regression of the estimated δ_{jt} on x_{jt} and p_{jt}.[55] As an example, Bayer, Ferreira,

[53] A similar observation limits the applicability of random utility models that represent all latent factors with a single (typically additive) random shock. For example, the random utility model in (6.1)–(6.3) could be rewritten more compactly as $u_{ijt} = \phi(x_{jt}, p_{jt}) + e_{ijt}$, allowing correlation between e_{ijt} and (x_{jt}, p_{jt}). But this representation alone is not adequate as a model of demand in such a setting. To measure an own-price demand elasticity, for example, one must hold ξ_{jt} (among other things) fixed while letting both the mean utility for good j and the distribution of the stochastic term $\mu_{ijt}(\nu_{it}; \beta_d, \beta_\nu)$ adjust to a change p_{jt}. By specifying only a composite stochastic term e_{ijt}, the more compact representation does not allow one to even define such *ceteris paribus* changes.

[54] Otherwise, the parameters δ may be poorly estimated.

[55] It will be necessary here to account for estimation error in the estimated δ_{jt}, so the correct standard errors for estimates constructed this way are not those of two-stage least squares.

and McMillan (2007) and Bayer and Timmins (2007) apply this two-step method to the demand for residential locations. Note that, in this parametric context, we now require just one excluded instrument, for the endogenous p_{jt}. In addition to cost shifters and cost proxies, candidate instruments include BLP instruments and Waldfogel instruments. In fact, unlike the case of market-level data, here one can use own-market forms of the Waldfogel instruments, such as the market-level means of d_{it}, as long as these are not elements of x_{jt} and not correlated with ξ_{jt}. Thus, the parametric micro data model requires as few as one excluded instrument (for price) to learn all parameters and, as compared to the market-level case, can allow the use of an additional class of instruments.

Although this two-step approach is instructive, it will often be preferable to estimate all parameters at once, exploiting both within-market and cross-market variation. At a minimum, this will typically aid efficiency. One-step estimation also avoids estimating the extra "parameters" δ—potentially a large number of them—when in fact these are defined in (6.3) as functions involving only α_0 and β_0 as unknown parameters.

A more subtle issue concerns identification. In the fully parametric model, it may indeed be possible to estimate the parameters (β_d, β_ν) using only within-market variation—even using data from only a single market.[56] But, the available results on nonparametric identification with micro data (see section 7) suggest that this possibility is dependent on the parametric structure. Indeed, the results there require both within-market and cross-market variation even to learn the relative effects of consumer observables d_{it} (the nonparametric analog of β_d here).

Exploiting all the variation in the data needed for nonparametric identification can often lead to much more precise estimates of parametric models. For example, Berry, Levinsohn, and Pakes (2004a) report that they tried to estimate a related random coefficients discrete choice model on micro data from a single market, but found that the estimates were very noisy. When they added choice-set variation (in the form of "second-choice" data—see section 6.3), the results became much more precise. This is consistent with the idea that some form of cross-market data may be important for learning about the parameters β_d and β_ν.

To estimate all parameters jointly, one could estimate using moment conditions reflecting the score of the likelihood (6.5) with respect to $(\delta, \beta_d, \beta_\nu)$ together with orthogonality conditions of the form

$$E\left[\left(\delta_{jt} - x_{jt}\beta_0 - \alpha_0 p_{jt}\right) z_{jt}\right] = 0,$$

where z_{jt} represents the exogenous x_{jt} combined with excluded instruments for p_{jt}. The latter take the same form as the orthogonality conditions used to estimate from market-level data, but without excluded instruments beyond those the for the price p_{jt}. Again, this illustrates a substantial advantage of micro data.

In practice estimation using a simulated likelihood function or simulated score function is sometimes unattractive. One reason is that simulating the log-likelihood (or its

[56] We are not aware of results characterizing the essential sources of variation for identification of the parametric model.

score) with sufficient precision can be computationally demanding, particularly when some true choice probabilities are close to zero. Train (2009) provides useful discussion and offers a number of computational tricks. Modern computational tools may expand the applicability of such approaches. However, a common alternative is to avoid the likelihood and rely instead on moment conditions capturing key variation across consumers and their choices.

Following Berry, Levinsohn, and Pakes (2004a), for example, one can combine moments reflecting market shares (typically fitting these exactly, as is usually done in the case of market-level data) with "micro moments" characterizing key features of the joint distribution of consumer i's characteristics and the characteristics of her choice $j(i)$. Typical micro moments include covariances, or conditional expectations of consumer characteristics given characteristics of the chosen product (or vice versa). As an example of the latter, in the case of autos one might use the conditional expectations of family size, age, and income conditional on the class of automobile (e.g., minivan, compact, luxury, pickup). This type of one-step method of moments approach again illustrates the more limited reliance on orthogonality conditions when one has micro data: in essence, aggregate moment conditions that pin down the nonlinear parameters in the case of market-level data can be replaced by micro moments that are sufficient to identify $(\delta, \beta_d, \beta_\nu)$.

The PyBLP software discussed previously provides code for estimating all parameters from micro data using certain combinations of market-level moments and "micro moments" like those discussed above. This open source code also provides a template for adapting the computation to incorporate estimation with other combinations of moments.[57]

6.2 Consumer Panels

Although what we call micro data is form of panel data, in the case of consumer demand the term "panel data" is typically reserved for something else: observation of each consumer on multiple choice occasions. For clarity, we refer to this as a "consumer panel." Examples include data on different grocery shopping trips for the same consumer, car purchases by the same family across different years, or health insurance selections by the same employee in the open enrollment period of different years. The advantage of a consumer panel is that observations on the same consumer on different choice occasions can provide even more information about the role of individual characteristics in determining substitution patterns. Intuitively, for example, one may directly observe which product a given consumer substitutes to in response to an exogenous price increase sufficient to induce a switch.

The estimation approaches discussed above are easily adapted to this setting. To illustrate, consider the same mixed logit model, but now splitting our usual notion of

[57] Although Conlon and Gortmaker (2020) does not cover micro data, documentation can be found at https://pyblp.readthedocs.io/en/stable/api.html#micro-moment-classes.

market into a geographically-defined market m and a time period t. We focus on the case in which one observes many consumers ("large N") on a small number of choice occasions ("small T") with many geographically-defined markets ("large M"). For simplicity, consider a two-period panel, so that $t \in \{0, 1\}$. We also focus on the case in which one views a consumer's random coefficients as reflecting stable preferences (thus, fixed across time periods), with only their product-specific shocks ϵ_{ijmt} drawn anew (and independently) each time period.

The model's prediction for the probability that a given consumer i in market m chooses good j in period 0 and good k in period 1 then takes the form (adapting the prior notation to the refined notion of markets)

$$s_{ijkm} = \int \left(\frac{e^{\delta_{jm0}+\mu_{ijm0}(\nu_{im};\beta_d,\beta_\nu)}}{1+\sum_\ell e^{\delta_{\ell m0}+\mu_{i\ell m0}(\nu_{im};\beta_d,\beta_\nu)}} \right) \left(\frac{e^{\delta_{km1}+\mu_{ikm1}(\nu_{im};\beta_d,\beta_\nu)}}{1+\sum_\ell e^{\delta_{\ell m1}+\mu_{i\ell m1}(\nu_{im};\beta_d,\beta_\nu)}} \right) dF_\nu(\nu_{im}). \tag{6.6}$$

Replacing j and k with the choices actually made by consumer i then yields the likelihood contribution for consumer i, as a function of the parameters $(\delta, \beta_d, \beta_\nu)$.[58] Just as in the case of micro data, the score of this likelihood function could be combined with additional moment conditions to yield an estimation strategy for all model parameters using the method of simulated moments and instruments for prices.[59]

The practical concerns associated with relying on a simulated log-likelihood with micro data can become more severe with panel data. Even with just two time periods, the choice probabilities above involve $(J+1)^2$ combinations of potential choices, and the true likelihood of some jk combinations observed in the data may, in certain contexts, be extremely small. Again, one could avoid such problems by instead using micro moments that incorporate some degree of aggregation. There are many possibilities, but a typical approach would start from the types of aggregate moments and micro moments discussed for estimation with micro data, adding moments capturing the extra information provided by the consumer panel: relationships between choices of the same consumer across choice occasions. For example, in the case of data on grocery purchases in a given product category, one might include as moments the covariances between the measured characteristics of products selected across shopping trips.

6.3 Ranked Choice Data

Some applications offer data on each consumer's rank ordering of products. Examples include certain types of conjoint analysis (common in marketing) or ranked school choices

[58] Implicit is an assumption of no state dependence—e.g., switching costs, rational inattention, or development of brand loyalty through past purchase. Allowing state dependence would require a change to the model, a choice of whether to treat consumers as forward-looking, and, typically, dealing with an "initial conditions" problem. See, e.g., Dubé, Hitsch, and Rossi (2010) and Handel (2013).

[59] Chintagunta and Dubé (2005) apply a similar strategy to a consumer panel of grocery store purchases. Similar to our discussion of the likelihood in (6.5), they estimate the mean utilities δ_{jt} and nonlinear parameters of the utility specification via maximum likelihood, then estimate the parameters of (6.3) with a second-step IV regression.

(e.g., within a strategy-proof school choice mechanism).[60] In principle, one could see consumer's ranking of all products, although typically one observes only a shortlist of top choices.

This type of ranked choice data can be thought of as an ideal form of a consumer panel. One observes, for each consumer, responses to the questions (a) what is your preferred product among all options? (b) what is your preferred product among all options excluding your favorite? (c) what is your preferred product among all those excluding your top two choices, ... This is very similar to a consumer panel offering observation of the same consumer's choice from multiple choice sets.[61] But there are at least two advantages to ranked choice data. First, the absence of temporal separation can avoid any question about which stochastic components of the model should be view as fixed across "choice occasions." Second, the type of "variation in the choice set" provided by ranked choice data is ideal for assessing which products are closest substitutes. Indeed, the substitution patterns that we have focused on as a primary challenge of demand estimation are closely connected to the relationships between first and second choices. Observing the first and second choices directly is therefore very powerful.[62] And, of course, the relationships among a longer list of ranked choices are driven by the same components of the model.

Estimation in the case of ranked choices can proceed along the lines suggested for a consumer panel. A likelihood approach could again be used, although with the same types of caveats.[63] The method of simulated moments again offers alternatives. Here, for example, one might combine market shares (average choice probabilities for first choices) with moments characterizing covariance between components of consumer characteristics, characteristics of first-choice products, and characteristics of first- and second-choice products (see Berry, Levinsohn, and Pakes (2004a)).

6.4 Hybrids

A common situation in practice is that one has a combination of multi-market market-level data and a limited set of micro data or ranked choice data. One prominent example

[60] Allenby, Hardt, and Rossi (2019) provides an overview of conjoint analysis. Hastings, Kane, and Staiger (2010) consider demand for ranked school choices and are followed by a large literature, including Agarwal and Somaini (2018).

[61] For at least some purposes, one will need to assume that the data provide correctly ranked preferences. In conjoint analysis, rankings are typically based on self-reports of hypothetical behavior, potentially introducing noise in the measured rankings. And in the case of school choice data, even with a strategy-proof school assignment mechanism, parents might not understand or trust the strategy-proof nature of the problem they face.

[62] See also the discussion in Conlon and Mortimer (2020).

[63] In the mixed logit model considered in this section, the joint probability of a consumer's rankings factors as a product of standard logit choice probabilities for appropriately defined choice sets, conditional on $(d_{it}, y_{it}, \nu_{it})$. See, e.g., Train (2009).

in the literature is Petrin's (2002) study of welfare gains from introduction of a new product, which combined aggregate market shares for automobiles with a small sample of micro-data from the Consumer Expenditure Survey (CEX). Another is Goeree's (2008) study of advertising and personal computer demand, which combined market shares for individual models with a limited micro-data sample linking individual characteristics to computer purchases by brand.[64] This kind of data configuration lends itself to the type of simulated method of moments estimation approach discussed above, combining (a) BLP-style moments involving aggregate market shares and exogeneity of instruments and (b) moments defined as the difference between predicted and actual (covariance or conditional mean) relationships between individual characteristics and characteristics of the product chosen.

7 Nonparametric Identification with Micro Data

Our discussion in the previous section suggests that micro data can be valuable both for allowing richer demand specifications and for estimation of the "nonlinear parameters" governing substitution patterns—adding precision and softening the reliance on instrumental variables. By returning to the question of nonparametric identification, we can demonstrate a strong formal confirmation of this message. In this section, we discuss recent results from Berry and Haile (2020) showing how the addition of micro data can allow a more flexible demand model and a reduction in both the number and types of instrumental variables required, as compared to the case of market-level data. Unsurprisingly, instruments for the endogenous prices will still be needed. However, micro data can eliminate the necessity of instruments that shift market shares independently of prices (cf. section 5.4.1). One can also allow for very flexible effects of consumer-level observables on demand—for example, avoiding any requirement that certain consumer observables alter preferences only for certain products.

Of course, replacing variation through relevant instruments with micro-data variation will require a notion of "sufficient" variation through the consumer observables, explained below. Furthermore, the identification of highly flexible effects of consumer observables on demand relies on a combination of within-market and cross-market variation: observed variation across consumers within a market does not suffice.

7.1 Nonparametric Demand Model

We will proceed somewhat less formally than in our discussion of nonparametric identification with market-level data (section 5), referring readers to Berry and Haile (2020) for a more complete treatment. Consider a nonparametric model of demand characterized

[64] Goeree (2008) also extends the BLP model to allow for incomplete "consideration sets," i.e., for consumers who are aware of only a strict subset of the available goods.

by equations

$$s_{ijt} = \sigma_j\left(d_{it}, y_{it}, x_t, p_t, \xi_t\right) \qquad j = 1, \ldots, J. \tag{7.1}$$

We will interpret s_{ijt} as the probability consumer i in market t chooses good j in a discrete choice demand model, although the arguments generalize to other settings (e.g., continuous demand, mixed discrete-continuous demand). As in our previous discussion of identification, we have conditioned on a given number of goods J. Compared to the model considered in the case of market-level data, here we have added observed individual-specific measures (d_{it}, y_{it}) as determinants of demand. The notation for the other arguments of σ_j is as in the preceding sections, with all goods' prices, product characteristics, and demand shocks entering the demand for each good j.

Here we will partition (d_{it}, y_{it}) so that $d_{it} \in \mathbb{R}^J$ and $y_{it} \in \mathbb{R}^H$, with $H \geq 0$. Thus, although there is no upper limit on the number of consumer-specific observables, we require at least as many consumer observables as goods. Furthermore, although we will require a nonparametric index restriction on the way d_{it} enters the demand model, we can accommodate other consumer observables y_{it} in an unrestricted way. The observables, for purposes of considering identification, consist of d_{it}, y_{it}, x_t, p_t, and choice probabilities conditional on (d_{it}, y_{it}) in each market t. In addition, there are observed excluded instruments that we discuss below. We treat the characteristics x_t as exogenous, as we have done in prior sections. To discuss identification we can then condition on an arbitrary value of x_t and suppress x_t in the notation below.[65]

In addition to the required degree of variation in d_{it}, choice sets and price instruments, the identification results in Berry and Haile (2020) rely on a set of core assumptions on demand. These play some of the roles of the "index and inversion" assumptions we discussed in the case of market-level data, although they take a different form. The four main assumptions are:

(i) For all j, $\sigma_j\left(d_{it}, y_{it}, p_t, \xi_t\right) = \sigma_j(\gamma(d_{it}, y_{it}, \xi_t), y_{it}, p_t)$, with $\gamma\left(d_{it}, y_{it}, \xi_t\right) \in \mathbb{R}^J$.

(ii) $\sigma\left(\cdot, y_{it}, p_t\right)$ is injective on the support of $\gamma(d_{it}, y_{it}, \xi_t)$ conditional on (y_{it}, p_t).

(iii) $\gamma\left(\cdot, y_{it}, \xi_t\right)$ is injective on the support of $d_{it}|y_{it}$

(iv) For all j, $\gamma_j\left(d_{it}, y_{it}, \xi_t\right) = g_j\left(d_{it}, y_{it}\right) + \xi_{jt}$.

Assumption (i) is a nonparametric index restriction. Like the index restriction in section 5, it limits the way that the demand shocks enter the model. We will connect this condition to standard specifications below. With the invertibility requirement of assumption (ii), this index restriction will ensure that we can construct an inverted demand system with only one structural error in each equation. A sufficient condition for assumption (ii) is the connected substitutes property discussed in section 5. Assumption

[65] Identification of demand conditional on *endogenous* x_t is an ongoing area of research, related to Berry and Haile (2020). Some models of endogenous x_t may allow one to condition on endogenous x_t and still identify demand responses to *ceteris paribus* changes in prices.

(iii) requires that the index vector itself be invertible with respect to the vector d_{it}. This is satisfied automatically in some standard specifications, such as linear random utility models in which each component d_{ijt} of d_{it} affects only the utility of good j.[66] Assumption (iii) avoids this requirement but maintains the requirement, given any value of the demand shocks ξ_t, of a one-to-one mapping between the vector d_{it} and the index vector. Finally, Assumption (iv) requires that the index function be additively separable. As with the identification results of section 5, a key motivation for this restriction is to allow use of the same instrument relevance condition that is required for identification in additively separable regression models.

To connect these assumptions to familiar parametric models, consider the mixed-logit random utility specification[67]

$$u_{ijt} = x_{jt}\beta_{ijt} - \alpha_{it}p_{jt} + \xi_{jt} + \epsilon_{ijt}, \tag{7.2}$$

where $\beta_{ijt}^{(k)} = \beta_{0j}^{(k)} + \sum_{\ell=1}^{L} \beta_{dj}^{(\ell,k)} d_{i\ell t} + \beta_{\nu j}^{(k)} \nu_{it}^{(k)}$ and $\ln(\alpha_{it}) = \alpha_0 + \alpha_y y_{it} + \alpha_\nu \nu_{it}^{(0)}$. We can rewrite (7.2) as

$$u_{ijt} = g_j\left(d_{it}, x_t\right) + \xi_{jt} + \mu_{ijt},$$

where

$$g_j\left(d_{it}, x_t\right) = \sum_k x_{jt}^{(k)} \sum_{\ell=1}^{L} \beta_{dj}^{(\ell,k)} d_{i\ell t} = \sum_{\ell=1}^{L} d_{i\ell t} \sum_k x_{jt}^{(k)} \beta_{dj}^{(\ell,k)}$$

and

$$\mu_{ijt} = \sum_k x_{jt}^{(k)} \left(\beta_{0j}^{(k)} + \beta_{\nu j}^{(k)} \nu_{it}^{(k)}\right) - p_{jt} \exp(\alpha_0 + \alpha_y y_{it} + \alpha_\nu \nu_{it}^{(0)}) + \epsilon_{ijt}$$

Thus, recalling that $L = J$,[68] our key assumptions hold (recall that those assumptions are stated conditional on the suppressed x_t, treating x_t fully flexibly) as long as the $J \times J$ matrix of coefficients on d_{it} (whose elements are $\sum_k x_{jt}^{(k)} \beta_{dj}^{(\ell,k)}$) is full rank.

[66] Such an exclusivity condition is often combined with assumptions of independence and "large support" to demonstrate identification through a "special regressor" argument (see, for example, Lewbel (2014)). The results discussed here will not require exclusivity, independence, or large support.

[67] This example generalizes the canonical model of section 3.2 by allowing random coefficients on x_{jt} to vary with j. With $L = J$, the common special case in which d_{ijt} enters only the utility for good j is then obtained by setting $\beta_{dj}^{(k,\ell)}$ to zero except when $k = 1$ (recall that the first component of each x_{jt} is a one) and $\ell = j$. Logit specifications allowing choice-specific coefficients on consumer characteristics are sometimes distinguished from models with choice-specific characteristics by labeling the former "multinomial logit" and the latter "conditional logit."

[68] Any additional observables are easily incorporated outside the mapping g without further restriction. The essential requirement is that consumer-level observables have dimension at least J.

7.2 Identification

Here we sketch the identification arguments, proceeding in three steps. First, a combination of within-market and cross-market variation is exploited to uncover the index function $g : \mathbb{R}^J \to \mathbb{R}^J$. Then cross-market variation—including that produced by excluded instruments for prices—allows identification of the demand shocks ξ_{jt} for all goods and markets in the same way that residuals in a nonparametric regression model are identified. Finally, with the demand shocks known, identification of demand is immediate from the definition of demand in (7.1). This argument does not require variation in y_{it}, so we henceforth fix y_{it} at an arbitrary value and suppress it in the notation.[69]

7.2.1 Identification of the Index Function

Let $\mathcal{S}(\xi, p)$ denote the support of the share vector when the random variables (ξ_t, p_t) take the values (ξ, p). Because d_{it} varies within each market, the set $\mathcal{S}(\xi, p)$ is not a singleton: each d_{it} in market t is associated with a different observed conditional choice probability vector s_{it}.

Given the assumptions on demand, for each vector of market shares $s \in \mathcal{S}(\xi, p)$ there will be a unique d^* in the support of d_{it} such that

$$\sigma(g(d^*) + \xi, p) = s.$$

This d^* is the vector of consumer characteristics that generate the choice probability vector s (given $(\xi_t, p_t) = (\xi, p)$). So we may write

$$d^*(s; \xi, p).$$

Furthermore, the inverted demand system at this point is

$$g(d^*(s; \xi, p)) + \xi = \sigma^{-1}(s; p). \tag{7.3}$$

Note that, because choice probabilities conditional on d_{it} in each market t are observed, $d^*(s; \xi_t, p_t)$ is observed for all t and $s \in \mathcal{S}(\xi_t, p_t)$ even though no ξ_t is observed or known at this point.

If we differentiate (7.3) within a market t where $p_t = p$ and $d^*(s; \xi_t, p) = d$, we obtain

$$\frac{\partial g(d)}{\partial d} \frac{\partial d^*(s; \xi_t, p)}{\partial s} = \frac{\partial \sigma^{-1}(s; p)}{\partial s}. \tag{7.4}$$

If we do the same within another market t' with the same p and same $s \in \mathcal{S}(\xi_{t'}, p)$, we get a similar expression with an identical right-hand side. Setting the two left-hand sides

[69] When y_{it} does vary, it can provide a source of overidentifying restrictions.

equal and letting $d' = d^*(s; \xi_{t'}, p)$, we see that

$$\frac{\partial g(d')}{\partial d} = \left[\frac{\partial g(d)}{\partial d}\right] \frac{\partial d^*(s; \xi_t, p)}{\partial s} \left[\frac{\partial d^*(s; \xi_{t'}, p)}{\partial s}\right]^{-1}. \tag{7.5}$$

The only unknowns in (7.5) are the matrices $\frac{\partial g(d')}{\partial d}$ and $\frac{\partial g(d)}{\partial d}$. Without loss, one may choose an arbitrary point d^0 and set both $g(d^0)$ and $\partial g(d^0)/\partial d$ to arbitrary values.[70] Thus, by starting at d^0 and stringing together relationships of the form (7.5) covering all points in the support of d_{it}, one can determine the derivatives of g on the entire support and integrate these derivatives up to the value of g at all such points.[71]

Observe that the function g determines the effects of d_{it} on demand relative to those of the demand shocks ξ_{jt}. Thus, identification of g is a key step toward identification of effects of d_{it} on demand. Because the latent ξ_t is fixed within a market and alters the demand of all consumers in that market, it may be unsurprising that identification of g would require both within-market and cross-market variation. Our proof in this step, indeed, uses both types of variation. Equations (7.3) and (7.4) make use of within-market variation, where ξ_t is fixed. Equation (7.5) then makes use of cross-market variation. Although $\frac{\partial d^*(s;\xi_t,p)}{\partial s}$ is observed in each market, equation (7.4) does not, by itself, allow one to distinguish $\frac{\partial g(d)}{\partial d}$ from $\partial \sigma^{-1}(s,p)/\partial s$. This indeterminacy is resolved by the cross-market variation used in equation (7.5).

7.2.2 Identification of Demand

If one hopes that micro data variation can substitute for some variation through instrumental variables, some requirement on the extent of variation in d_{it} will of course be required. Berry and Haile (2020) require that there exist some "common choice probability" vector s^* that is reached in every market by a consumer with the "right" characteristics d_{it} for that market. Specifically, to our earlier assumptions (i)–(iv) we add

(v) There exists s^* such that $s^* \in \mathcal{S}(\xi, p)$ for all $(\xi, p) \in \text{supp}\,(\xi_t, p_t)$.

As an example, with two inside goods and outside good this assumption requires existence of at least one vector of choice probabilities for the inside goods—perhaps 0.4 for each—that is reached in every market t by conditioning on the right type of consumer (i.e., the right vector of observables (d_{1t}, d_{2t})) for that market. We need not know in advance what this common choice probability vector s^* is. Indeed, whether such a vector s^* exists is observable—formally, this assumption is "verifiable" (see, e.g., Berry and Haile (2018)).

The strength of the common choice probability requirement in a given application will depend on the supports of p_t and ξ_t and the effects that each of these has on demand.

[70] As discussed by Berry and Haile (2020), these are proper normalizations imposing no restriction on the demand system (7.1).

[71] Berry and Haile (2020) provide technical conditions ensuring that this is possible.

A helpful fact is that higher values of p_{jt} and ξ_{jt} typically have opposite effects, whereas equilibrium pricing typically implies positive dependence between p_{jt} and ξ_{jt}. Thus, the effects of even large variation in the demand shocks may be substantially damped by the accompanying variation in prices. In any case, this assumption strictly relaxes the "large support" assumption sometimes relied upon to ensure nonparametric identification in discrete choice models. A large support assumption on d_{it} would require sufficient variation to move choice probabilities to every point in the simplex in every market; i.e., large support would imply that every vector of nonzero choice probabilities is a common choice probability. Here only one such vector is required.

With a common choice probability vector s^*, in every market t we have J inverse demand equations of the form

$$g_j\left(d^*\left(s^*;\xi_t,p_t\right)\right)=\sigma_j^{-1}\left(s^*;p_t\right)-\xi_{jt}. \tag{7.6}$$

In each equation, the left-hand side is now known. On right-hand side, s^* is fixed across markets. Thus, each equation (7.6) takes the form of a nonparametric regression equation with a separable structural error. Identification of the "regression function" $\sigma_j^{-1}\left(s^*;p_t\right)$ then follows immediately from the identification result of Newey and Powell (2003) given instruments for the endogenous variables—p_t here—satisfying the standard mean independence (exclusion) and completeness (relevance) conditions.

This immediately implies identification of each ξ_{jt} as well. With the demand shocks ξ_{jt} known, identification of demand follows immediately from equation (7.1), since choice probabilities s_{ijt} are observed and all arguments of the functions $\sigma_j(\cdot)$ are now known.

7.3 Discussion

These results provide formal confirmation of a key benefit of micro data suggested in our discussion of the parametric models: micro data with sufficient variation, combined with choice sets that vary appropriately across markets, allow us to replace instruments for quantities with micro-data variation via d_{it}. The need for J-dimensional variation in d_{it} is directly connected to the J endogenous quantities. Note, however, that the "exogeneity" of the micro-data variation arises not from an exclusion restriction but from the fact that within a single market, market-level demand shocks simply do not vary. Thus, with micro data, one has many of the same advantages that allow "within estimation" of slope parameters in other types of panel data models.

These results imply that micro data can cut the number of required instruments by half. Related but distinct is the fact that the BLP instruments are no longer required. A benefit of the latter is our ability to treat all components of x_t in a fully flexible way, considering a strictly more general model than that considered in section 5 for the case of market-level data. Of course, as discussed in section 5.4.3, this fully general treatment of x_t also implies that BLP instruments are not even available as candidate instruments. BLP instruments can be made available by requiring some components of x_t to enter through the indices γ_t (see Berry and Haile (2020) for details). And the other types of instruments discussed in section 4.2 remain valid candidates to instrument for prices here.

Finally, although the formal results address only the case in which micro data variation fully replaces J instruments, we expect that micro data with more limited variation will still be useful in practice. There are many applied cases where one has only "partial" micro data, or micro data with more limited dimension. In these cases, an applied researcher might employ a combination of BLP-style instruments and moment conditions reflecting the available micro data, as discussed in section 6.

8 Some Directions for Future Work

We have focused on "foundations" of demand estimation—the key challenges, core models, standard methods, and sources of identification. This focus necessarily leaves many topics neglected or considered only very lightly. A number of these topics offer fruitful directions for future research.

One is dynamics. Many demand decisions are dynamic, leading to potentially important roles for various kinds of sunk costs and adjustment costs that interact with tastes that are persistent through time. Models incorporating such factors raise difficult computational issues, and also difficult questions about how to discriminate between (or separately identify) various forms of state dependence and persistent tastes. Interesting reading here includes Hendel and Nevo (2006), Gowrisankaran and Rysman (2012), and Handel (2013), as well as the enormous literature on dynamic models more generally, where distinguishing between unobserved heterogeneity and state dependence has long been a focus (see, e.g., Heckman (1981)).

A second topic concerns functional forms used for estimation. We noted that the widely-used mixed logit model offers a compromise between flexibility and the parsimony necessary for real-world estimation. This compromise was designed in particular to provide flexibility (at least as compared to a multinomial logit or CES model) in the cross-product derivatives with respect to prices and product characteristics. These substitution patterns drive answers to many questions of interest—e.g., the sizes of markups or outcomes under a counterfactual merger. However, other kinds of counterfactuals can require flexibility in other dimensions. For example, "pass-through" (e.g., of a tariff, tax, or technologically driven reduction in marginal cost) depends critically on second-derivatives of demand. It is not clear that a mixed-logit model is very flexible in this dimension. An alternative is nonparametric demand estimation, as in Compiani (2020), although many off-the-shelf nonparametric approaches lack the parsimony necessary to estimate demand systems with a large number of products or product characteristics. An interesting question is whether alternative (parametric, semi-parametric, or nonparametric) specifications of demand or the distribution of utilities can offer attractive alternatives.[72]

[72] See, e.g., Gandhi, Nevo, and Tao (2019b).

A third topic of interest is invertibility of demand. Inversion of the demand system arises repeatedly our discussions of both identification and estimation. This reflects a reliance on inversion to address the fundamental challenges of simultaneity and the appearance of many structural errors (demand shocks) in each demand equation. But invertibility of a demand system is not automatic, and demand systems often violate standard conditions used to establish invertibility (i.e., injectivity or "univalence") of multivariate maps. Berry, Gandhi, and Haile (2013) discuss this issue and offer invertibility conditions that can accommodate some forms of complementarity. However, invertibility (or alternatives) in the case of complements has not been fully explored.[73] The same is true of invertibility in models in which consumers make multiple purchases. A related question is the extent to which there are helpful estimation approaches—perhaps involving partial identification—for settings in which invertibility fails.

The necessity of further progress should not obscure progress made, however. Compared to the situation 10 or so years ago, we now have access to a set of useful identification results, greater clarity about the role of instruments and the types of instruments available for estimating demand, more fully developed asymptotic theory, and set of now well-proven computational tools for constructing standard estimators. This progress suggests that much of the best future research will involve serious application of these existing tools.

[73] For a simple approach allowing a particular kind of complementary, see also Fosgerau, Monardo, and De Palma (2020).

References

Ackerberg, D. (2003): "Advertising Learning and Customer Choice in Experience Good Markets: A Structural Empirical Examination," *International Economic Review*, 44, 1007–1040.

Adao, R., A. Costinot, and D. Donaldson (2017): "Nonparametric Counterfactual Predictions in Neoclassical Models of International Trade," *American Economic Review*, 107, 633–89.

Agarwal, N. and P. Somaini (2018): "Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism," *Econometrica*, 86, 391–444.

Allenby, G. M., N. Hardt, and P. E. Rossi (2019): "Chapter 3 - Economic foundations of conjoint analysis," in *Handbook of the Economics of Marketing, Volume 1*, ed. by J.-P. Dub and P. E. Rossi, North-Holland, vol. 1 of *Handbook of the Economics of Marketing*, 151 – 192.

Anderson, S., A. DePalma, and F. Thisse (1992): *Discrete Choice Theory of Product Differentiation*, Cambridge MA: MIT Press.

Andrews, D. W. K. and P. Guggenberger (2017): "Asymptotic Size of Kleibergens LM and conditional LR Tests for Moment Condition Models," *Econometric Theory*, 33, 10461080.

Andrews, I. (2018): "Valid Two-Step Identification-Robust Confidence Sets for GMM," *Review of Economics and Statistics*, 100, 337–348.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quarterly Journal of Economics*, 132, 1553–1592.

Andrews, I. and A. Mikusheva (2020): "Optimal Decision Rules for Weak GMM," Working paper, Harvard.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 67, pp. 499–527.

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

Armstrong, T. B. (2016): "Large Market Asymptotics for Differentiated Product Demand Estimators with Economic Models of Supply," *Econometrica*, 84, 1961–1980.

Backus, M., C. Conlon, and M. Sinkinson (2020): "Common Ownership and Competition in the Ready-to-Eat Cereal Industry," Working paper, New York University.

Bayer, P., F. Ferreira, and R. McMillan (2007): "A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of Political Economy*, 115, 588–638.

Bayer, P. and C. Timmins (2007): "Estimating Equilibrium Models of Sorting Across Locations," *The Economic Journal*, 117, 353–374.

Ben-Akiva, M. E. (1973): "Structure of Passenger Travel Demand Models," Ph.D. thesis, MIT Department of Civi Engineering.

Benkard, L. and S. T. Berry (2006): "On the Nonparametric Identification of Nonlinear Simultaneous Equations Models: Comment on Brown (1983) and Roehrig (1988)," *Econometrica*, 74, 1429–1440.

Berry, S. (1994): "Estimating Discrete Choice Models of Product Differentiation," *RAND Journal of Economics*, 23, 242–262.

Berry, S., M. Carnall, and P. Spiller (1996): "Airline Hubs: Costs, Markups and the Implications of Consumer Heterogeneity," Working paper no. 5561, NBER.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 60, 889–917.

——— (1999): "Voluntary Export Restraints on Automobiles: Evaluating a Strategic Trade Policy," *American Economic Review*, 89, 189–211.

——— (2004a): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Vehicle Market," *Journal of Political Economy*, 112, 68–105.

Berry, S., O. Linton, and A. Pakes (2004b): "Limit Theorems for Differentiated Product Demand Systems," *Review of Economic Studies*, 71, 613–614.

Berry, S. T., A. Gandhi, and P. A. Haile (2013): "Connected Substitutes and Invertibility of Demand," *Econometrica*, 81, 2087–2111.

Berry, S. T. and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

——— (2018): "Nonparametric Identification of Simultaneous Equations Models with a Residual Index Structure," *Econometrica*, 86, 289–315.

——— (2020): "Nonparametric Identification of Differentiated Products Demand Using Micro Data," Working paper no. 27704, National Bureau of Economic Research.

Bhattacharya, D. (2018): "Empirical Welfare Analysis for Discrete Choice: Some General Results," *Quantitative Economics*, 9, 571–615.

Blundell, R., D. Kristensen, and R. L. Matzkin (2017): "Individual Counterfactuals with Multidimensional Unobserved Heterogeneity," Tech. rep., CEMMaP Working Paper CWP60/17.

Blundell, R. and R. L. Matzkin (2014): "Control Functions in Nonseparable Simultaneous Equations Models," *Quantitative Economics*, 5, 271–295.

Bresnahan, T. (1981): "Departures from Marginal Cost Pricing in the American Automobile Industry," *Journal of Econometrics*, 17, 201–227.

——— (1987): "Competition and Collusion in the American Automobile Oligopoly: The 1955 Price War," *Journal of Industrial Economics*, 35, 457–482.

Brown, B. (1983): "The Identification Problem in Systems Nonlinear in the Variables," *Econometrica*, 51, 175–96.

Burda, M., M. Harding, and J. Hausman (2015): "A Bayesian Mixed Logit-Probit Model for Multinomial Choice," *Journal of Applied Econometrics*, 30, 353–376.

Capps, C., D. Dranove, and M. Satterthwaite (2003): "Competition and Market Power in Option Demand Markets," *RAND Journal of Economics*, 34, 737–763.

Chamberlain, G. (1986): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 305–334.

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

Chintagunta, P. K. and J.-P. Dubé (2005): "Estimating a Stockkeeping-Unit-Level Brand Choice Model that Combines Household Panel Data and Store Data," *Journal of Marketing Research*, 42, 368–379.

Compiani, G. (2020): "Market Counterfactuals and the Specification of Multi-Product Demand: A Nonparametric Approach," Working paper, University of Chicago.

Conlon, C. and J. Gortmaker (2020): "Best Practices for Differentiated Products Demand Estimation with pyBLP," *RAND Journal of Economics*, 51, 1108–1161.

Conlon, C. and J. Mortimer (2020): "Empirical Properties of Diversion Ratios," working paper, NYU.

Debreu, G. (1960): "Review of *Individual Choice Behavior: A Theoretical Analysis* by R. Duncan Luce," *American Economic Review*, 50, 186–188.

DellaVigna, S. and M. Gentzkow (2019): "Uniform Pricing in U.S. Retail Chains*," *The Quarterly Journal of Economics*, 134, 2011–2084.

Diamond, R. (2016): "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000," *The American Economic Review*, 106, 479–524.

Duarte, M., L. Magnolfi, M. Sølvsten, and C. Sullivan (2021): "Testing Firm Conduct," working paper, University of Wisconsin-Madison.

Dubé, J.-P., J. T. Fox, and C.-L. Su (2012): "Improving the Numerical Performance of Static and Dynamic Aggregate Discrete Choice Random Coefficients Demand Estimation," *Econometrica*, 80, pp. 2231–2267.

Dubé, J.-P., G. Hitsch, and P. Rossi (2010): "State Dependence and Alternative Explanations for Consumer Inertia," *RAND Journal of Economics*, 41, 417–445.

Eizenberg, A. (2014): "Upstream Innovation and Product Variety in the United States Home PC Market," *Review of Economic Studies*, 81, 1003–1045.

Fan, Y. (2013): "Ownership Consolidation and Product Characteristics: A Study of the U.S. Daily Newspaper Market," *American Economic Review*, 103, 1598–1628.

Fosgerau, M., J. Monardo, and A. De Palma (2020): "The Inverse Product Differentiation Logit Model," Working paper, CREST.

Freyberger, J. (2015): "Asymptotic theory for differentiated products demand models with many markets," *Journal of Econometrics*, 185, 162 – 181.

Gandhi, A. and J.-F. Houde (2020): "Measuring Substitution Patterns in Differentiated Products Industries," Working paper, Univ of Wisconsin.

Gandhi, A., Z. Lu, and X. Shi (2019a): "Estimating Demand for Differentiated Products with Zeroes in Market Share Data," Working paper, University of Wisconsin-Madison.

Gandhi, A., A. Nevo, and J. Tao (2019b): "Flexible Estimation of Differentiated Products Demand Models Using Aggregate Data," Tech. rep., University of Pennsylvania.

Gentzkow, M. (2007): "Valuing New Goods in a Model with Complementarities: Online Newspapers," *American Economic Review*, 97, 713–744.

Gentzkow, M. and J. Shapiro (2010): "What Drives Media Slant? Evidence from U.S. Newspapers," *Econometrica*, 78, 35–71.

Gillen, B. J., H. R. Moon, S. Montero, and M. Shum (2019): "BLP2-Lasso for Aggregate Discrete Choice Models with Rich Covariates," *The Econometrics Journal*, 22, 262–281.

Gillen, B. J., H. R. Moon, and M. Shum (2014): "Demand Estimation with High-dimensional Product Characteristics," in *Advances in Econometrics: Bayesian Model Comparison*, ed. by I. Jeliazkov and D. Poirier, Emerald Publishing, vol. 34.

Goeree, M. S. (2008): "Limited Information and Advertising in the US Personal Computer Industry," *Econometrica*, 76, 1017–1074.

Goldberg, P. K. (1995): "Product Differentiation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry," *Econometrica*, 63, 891–951.

Gowrisankaran, G. and M. Rysman (2012): "Dynamics of Consumer Demand for New Durable Goods," *JPE*, 120, 1173–1219.

Handel, B. (2013): "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *American Economic Review*.

Hastings, J., T. Kane, and D. Staiger (2010): "Heterogeneous Preferences and the Efficacy of Public School Choice," Tech. rep., Brown University.

Hausman, J., G. Leonard, and J. Zona (1994): "Competitive Analysis with Differentiated Products," *Annales d'Economie et de Statistique*, 34, 159–180.

Hausman, J. and D. Wise (1978): "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46, 403–426.

Hausman, J. A. (1996): "Valuation of New Goods under Perfect and Imperfect Competition," in *The Economics of New Goods*, ed. by T. F. Bresnahan and R. J. Gordon, Chicago: University of Chicago Press, chap. 5, 209–248.

Heckman, J. J. (1981): "Heterogeneity and State Dependence," in *Studies in Labor Markets*, ed. by S. Rosen, University of Chicago Press.

Hendel, I. and A. Nevo (2006): "Sales and consumer inventory," *The RAND Journal of Economics*, 37, 543–561.

Ho, K. (2009): "Insurer-Provider Networks in the Medical Care Market," *American Economic Review*, 99, 393–430.

Hom, M. (2018): "School Choice, Segregation and Access to Quality Schools: Evidence from Arizona," Working paper, Yale University.

Hong, H., H. Li, and J. Li (2020): "BLP estimation using Laplace transformation and overlapping simulation draws," *Journal of Econometrics*.

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512.

Kim, K. and A. Petrin (forthcoming): "Control Function Corrections for Unobserved Factors in Differentiated Products Models," *Quantitative Marketing and Economics.*

Knittel, C. R. and K. Metaxoglou (2014): "Estimation of Random-Coefficient Demand Models: Two Empiricists' Perspective," *Review of Economics and Statistics*, 96, 34–59.

Koopmans, T. C. (1949): "Identification Problems in Economic Model Construction," *Econometrica*, 17, 125–144.

Lewbel, A. (2014): "An Overview of the Special Regressor Method," in *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, ed. by J. S. Racine, L. Su, and A. Ullah, Oxford University Press, 38–62.

MacKay, A. and N. H. Miller (2021): "Estimating Models of Supply and Demand: Instruments and Covariance Restrictions," Tech. rep., Harvard.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995): *Microeconomic Theory*, Oxford University Press.

Matzkin, R. L. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375.

——— (2008): "Identification in Nonparametric Simultaneous Equations," *Econometrica*, 76, 945–978.

——— (2015): "Estimation of Nonparametric Models with Simultaneity," *Econometrica*, 83, 1–66.

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers of Econometrics*, ed. by P. Zarembka, New York: Academic Press.

——— (1978): "Modelling the Choice of Residential Location," in *Spatial Interaction Theory and Planning Models*, ed. by A. Karlvist, Amsterdam: North Holland, 75–96.

——— (1981): "Econometric Models of Probabilistic Choice," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden, Cambridge, MA: MIT Press.

McFadden, D., A. Talvitie, and Associates (1977): *Demand Model Estimation and Validation*, Berkeley CA: Institute of Transportation Studies.

Miller, N. H. and M. C. Weinberg (2017): "Understanding the Price Effects of the MillerCoors Joint Venture," *Econometrica*, 85, 1763–1791.

Neilson, C. (2020): "Targeted Vouchers, Competition Among Schools, and the Academic Achievement of Poor Students," Working paper, Princeton University.

Nevo, A. (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69, 307–42.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation in Nonparametric Models," *Econometrica*, 71, 1565–1578.

Petrin, A. (2002): "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 110, 705–729.

Petrin, A. and K. Train (2010): "A Control Function Approach to Endogeneity in Consumer Choice Models," *Journal of Marketing Research*, 47, 3–13.

Quan, T. W. and K. R. Williams (2018): "Product Variety, Across Market Demand Heterogeneity, and the Value of Online Retail," *The RAND Journal of Economics*, 49, 877–913.

Quandt, R. E. (1956): "A Probabilistic Theory of Consumer Behavior," *Quarterly Journal of Economics*, 70, 507–536.

——— (1966): "A Probabilistic Abstract Mode Model," in *Studies in Travel Demand, Volume II*, Mathematica, Princeton, N.J., 90–113.

——— (1968): "Estimation of Modal Splits," *Transportation Research*, 2, 41–50.

Reynaert, M. and F. Verboven (2014): "Improving the performance of random coefficients demand models: The role of optimal instruments," *Journal of Econometrics*, 179, 83 – 98.

Rosse, J. N. (1970): "Estimating Cost Function Parameters without using Cost Function Data: An Illustrated Methodology," *Econometrica*, 38, 256–275.

Salanie, B. and F. A. Wolak (2019): "Fast, 'Robust', and Approximately Correct: Estimating Mixed Demand Systems," Working Paper 25726, National Bureau of Economic Research.

Train, K. E. (2009): *Discrete Choice Methods with Simulation*, Cambridge Press, 2nd ed.

Waldfogel, J. (2003): "Preference Externalities: An Empirical Study of Who Benefits Whom in Differentiated-Product Markets," *RAND Journal of Economics*, 34, 557–568.

Williams, K. R. and B. M. Adams (2019): "Zone Pricing in Retail Oligopoly," *American Economic Journal: Microeconomics*, 11, 124–156.