Chapter 64

STRUCTURAL ECONOMETRIC MODELING: RATIONALES AND EXAMPLES FROM INDUSTRIAL ORGANIZATION

PETER C. REISS

Graduate School of Business, Stanford University, Stanford, CA 94305-5015, USA
e-mail: preiss@optimum.stanford.edu

FRANK A. WOLAK

Department of Economics, Stanford University, Stanford, CA 94305-6072, USA
e-mail: wolak@zia.stanford.edu

Contents

Abstract 4280
Keywords 4280
1. Introduction 4281
2. Structural models defined 4282
3. Constructing structural models 4285
3.1. Sources of structure 4285
3.2. Why add structure? 4288
3.3. Evaluating structure – single equation models 4290
3.4. Evaluating structure – simultaneous equation models 4293
3.5. The role of nonexperimental data in structural modeling 4301
4. A framework for structural econometric models in IO 4303
4.1. The economic model 4304
4.2. The stochastic model 4304
4.2.1. Unobserved heterogeneity and agent uncertainty 4305
4.2.2. Optimization errors 4308
4.2.3. Measurement error 4311
4.3. Steps to estimation 4312
4.4. Structural model epilogue 4314
5. Demand and cost function estimation under imperfect competition 4315
5.1. Using price and quantity data to diagnose collusion 4315
5.2. The economic model 4317
5.2.1. Environment and primitives 4317

Handbook of Econometrics, Volume 6A

DOI: 10.1016/S1573-4412(07)06064-3

5.2.2. Behavior and optimization 4318
5.2.3. The stochastic model 4320
5.3. Summary 4324
6. Market power models more generally 4325
6.1. Estimating price–cost margins 4326
6.2. Identifying and interpreting price–cost margins 4329
6.3. Summary 4333
7. Models of differentiated product competition 4334
7.1. Neoclassical demand models 4334
7.2. Micro-data models 4340
7.2.1. A household-level demand model 4342
7.2.2. Goldberg's economic model 4342
7.2.3. The stochastic model 4344
7.2.4. Results 4347
7.3. A product-level demand model 4348
7.3.1. The economic model in BLP 4349
7.3.2. The stochastic model 4350
7.4. More on the econometric assumptions 4353
7.4.1. Functional form assumptions for price 4353
7.4.2. Distribution of consumer heterogeneity 4355
7.4.3. Unobserved "product quality" 4357
7.4.4. Cost function specifications 4359
7.5. Summary 4360
8. Games with incomplete information: Auctions 4361
8.1. Auctions overview 4361
8.1.1. Descriptive models 4363
8.1.2. Structural models 4365
8.1.3. Nonparametric identification and estimation 4368
8.2. Further issues 4374
8.3. Parametric specifications for auction market equilibria 4375
8.4. Why estimate a structural auction model? 4379
8.5. Extensions of basic auctions models 4381
9. Games with incomplete information: Principal-agent contracting models 4382
9.1. Observables and unobservables 4383
9.2. Economic models of regulator–utility interactions 4385
9.3. Estimating productions functions accounting for private information 4388
9.3.1. Symmetric information model 4391
9.3.2. Asymmetric information model 4391
9.4. Econometric model 4393
9.5. Estimation results 4397
9.6. Further extensions 4398
10. Market structure and firm turnover 4398
10.1. Overview of the issues 4399

10.1.1. Airline competition and entry 4400
10.2. An economic model and data 4402
10.3. Modeling profits and competition 4404
10.4. The econometric model 4406
10.5. Estimation 4409
10.6. Epilogue 4410
11. Ending remarks 4411
References 4412

Abstract

This chapter explains the logic of structural econometric models and compares them to other types of econometric models. We provide a framework researchers can use to develop and evaluate structural econometric models. This framework pays particular attention to describing different sources of unobservables in structural models. We use our framework to evaluate several literatures in industrial organization economics, including the literatures dealing with market power, product differentiation, auctions, regulation and entry.

Keywords

structural econometric model, market power, auctions, regulation, entry

JEL classification: C50, C51, C52, D10, D20, D40

1. Introduction

The founding members of the Cowles Commission defined *econometrics* as: "a branch of economics in which economic theory and statistical method are fused in the analysis of numerical and institutional data" [Hood and Koopmans (1953, p. xv)]. Today economists refer to models that combine explicit economic theories with statistical models as *structural econometric models*.

This chapter has three main goals. The first is to explain the logic of structural econometric modeling. While structural econometric models have the logical advantage of detailing the economic and statistical assumptions required to estimate economic quantities, the fact that they impose structure does not automatically make them sensible. To be convincing, structural models minimally must be: (1) flexible statistical descriptions of data; (2) respectful of the economic institutions under consideration; and, (3) sensitive to the nonexperimental nature of economic data. When, for example, there is little economic theory on which to build, the empiricist may instead prefer to use nonstructural or descriptive econometric models. Alternatively, if there is a large body of relevant economic theory, then there may significant benefits to estimating a structural econometric model – provided the model can satisfy the above demands.

A second goal of this chapter is to describe the ingredients of structural models and how structural modelers go about evaluating them. Our discussion emphasizes that the process of building a structural model involves a series of related steps. These steps are by no means formulaic and often involve economic, statistical and practical compromises. Understanding when and why structural modelers must make compromises, and that structural modelers can disagree on compromises, is important for understanding that structural modeling is in part "art". For example, structural modelers often introduce "conditioning variables" that are not explicitly part of the economic theory as a way of controlling for plausible differences across observations.

Our third goal is to illustrate how structural modeling tradeoffs are made in practice. Specifically, we examine different types of structural econometric models developed by industrial organization ("IO") economists. These models examine such issues as: the extent of market power possessed by firms; the efficiency of alternative market allocation mechanisms (e.g., different rules for running single and multi-unit auctions); and the empirical implications of information and game-theoretic models. We should emphasize that this chapter is NOT a comprehensive survey of the IO literature or even a complete discussion of any single topic. Readers interested in a comprehensive review of a particular literature should instead begin with the surveys we cite. Our goal is instead to illustrate selectively how IO researchers have used economic and statistical assumptions to identify and estimate economic magnitudes. Our hope is that in doing so, we can provide a better sense of the benefits and limitations of structural econometric models.

We begin by defining structural econometric models and discussing when one would want to use a structural model. As part of this discussion, we provide a framework

for evaluating the benefits and limitations of structural models. The remainder of the chapter illustrates some of the practical tradeoffs IO researchers have made.

2. Structural models defined

In structural econometric models, economic theory is used to develop mathematical statements about how a set of observable "endogenous" variables, y, are related to another set of observable "explanatory" variables, x. Economic theory also may relate the y variables to a set of unobservable variables, ξ. These theoretical relations usually are in the form of equalities: $y = g(x, \xi, \Theta)$, where $g(\cdot)$ is a known function and Θ a set of unknown parameters or functions. Occasionally, economic theory may only deliver inequality relations, such as $y \geqslant g(x, \xi, \Theta)$.

Economic theory alone usually does not provide enough information for the econometrician to estimate Θ. For this reason, and because the economic model $y = g(x, \xi, \Theta)$ may not be able to rationalize the observed data perfectly, the econometrician adds statistical assumptions about the joint distribution of x, ξ and other unobservables appended to the model. Taken together, these economic and statistical assumptions define an empirical model that is capable of rationalizing all possible observable outcomes. In order to estimate the underlying primitives of this model, researchers use statistical objects based on the model, such as a log-likelihood function for the data, $\ell(y, x \mid \Theta)$, or conditional moments, such as $E(y \mid x, \Theta)$.

Nonstructural empirical work in economics may or may not be based on formal statistical models. At one end of the spectrum are measurement studies that focus on constructing and summarizing data, such as labor force participation and unemployment rates. At the other end are those that use formal statistical models, such as autoregressive conditional volatility models. Both types of nonstructural empirical work have a long and respected tradition in economics. An excellent early example is Engel's (1857) work relating commodity budget shares to total income. Engel's finding that expenditure shares for food were negatively related to the logarithm of total household expenditures has shaped subsequent theoretical and empirical work on household consumption behavior [see Deaton and Muellbauer (1980) and Pollak and Wales (1992)]. A somewhat more recent example of descriptive work is the Phillips curve. Phillips (1958) documented an inverse relationship between United Kingdom unemployment rates and changes in wage rates. This work inspired others to document relationships between unemployment rates and changes in prices. In the ensuing years, many economic theories have been advanced to explain why Phillips curves are or are not stable economic relations.

Nonstructural empirical models usually are grounded in economics to the extent that economics helps identify which variables belong in y and which in x. This approach, however, ultimately estimates characteristics of the joint population density of x and y, $f(x, y)$, or objects that can be derived from it, such as:

$f(y \mid x)$, the conditional density of y given x;
$E(y \mid x)$, the conditional expectation of y given x;
$\mathrm{Cov}(y \mid x)$, the conditional covariances (or correlations) of y given x; or,
$Q_\alpha(y \mid x)$ the α conditional quantile of y given x.

The most commonly estimated characteristic of the joint density is the best linear predictor (BLP($y \mid x$)) of y given x.

More recently researchers have taken advantage of developments in nonparametric and semiparametric statistical methods to derive consistent estimates of the joint density of y and x. For example, statisticians have proposed kernel density techniques and other data smoothing methods for estimating $f(x, y)$. These same smoothing techniques have been used to develop nonparametric conditional mean models. Silverman (1986), Härdle (1990), Härdle and Linton (1994) and others provide useful introductions to these procedures. A major advantage of nonparametric models is that they can provide flexible descriptions of the above statistical quantities.

Given their flexibility, it would seem that nonstructural empirical researchers should always prefer nonparametric methods. There are, however, limitations to nonparametric methods. One is that they may require large amounts of data to yield much precision.[1] Second, and more important, once estimated, it is unclear how a flexible estimate of a joint density can be used to recover economic constructs such as economies of scale in production and consumer welfare. Moreover, it is also unclear how to perform out-of-sample counterfactual calculations, such as the impact of an additional bidder on the winning bid in an auction.

It is tempting to look at our descriptions of structural versus nonstructural models, and parametric versus nonparametric models, and see them as absolutes – empirical models are either structural or nonstructural, parametric or nonparametric. We see little value in such absolute classification exercises. In practice, it is not uncommon to find structural econometric models that include nonstructural components or nonparametric components. Our goal in providing these definitions is to have an initial basis for classifying and evaluating the success of an econometric model.

An initial example from IO may help understand our focus and intent. Consider a researcher who observes the winning bids, $y = \{y_1, \ldots, y_T\}'$, in each of a large number of T similar auctions. Suppose the researcher also observes the number of bidders, $x = \{x_1, \ldots, x_T\}'$, in each auction. To understand the equilibrium relationship between winning bids and the number of bidders the researcher could use a structural or a nonstructural approach.

[1] Silverman (1986) argues that researchers using these techniques face a "curse of dimensionality", wherein the amount of data required to obtain precise estimates grows rapidly with the dimensions of x and y. His calculations (1986, Table 4.2) suggest that researchers may need thousands, if not hundreds of thousands of observations before they can place great faith in these flexible techniques. For instance, more than ten times as much data is required to attain the same level of precision for a four-dimensional as a two-dimensional joint density. More than 200 times as much data is required for an eight-dimensional as a four-dimensional density.

A standard nonstructural approach would be to regress winning bids on the number of bidders. Under standard statistical assumptions, this regression would deliver the best linear predictor of winning bids given the number of bidders. These coefficient estimates could be used to predict future winning bids as a function of the number of bidders. Alternatively, a researcher worried about a nonlinear relationship between winning bids and the number of bidders might instead opt to use nonparametric smoothing techniques to estimate the conditional density of winning bids given each distinct observed number of bidders x, $f(y \mid x)$. The researcher could then use this estimated conditional density, $\widehat{f(y \mid x)}$, to calculate whether, for example, expected winning bids in the sample auctions increased or decreased with the number of bidders. The researcher could also check to see if the conditional expected bid increased or decreased linearly with the number of bidders.

The process of formulating and implementing either of these nonstructural models so far has made little use of economics, except perhaps to identify what is y and what is x. For instance, our discussion of these descriptive models has made no reference to institutional features of the auctions (e.g., sealed-bid versus open-outcry and first-price versus second-price). It also has not required economic assumptions about bidder behavior or characteristics (e.g., risk-neutrality, expected profit maximization and bidder competition). In some cases (e.g., sealed-bid versus open-outcry), we may be able to incorporate these considerations into a nonstructural model by introducing them as conditioning variables. In other cases (e.g., the degree of risk aversion), this may not be possible.

A key reason then to use economic theory, beyond specifying x and y, is to clarify how institutional and economic conditions affect relationships between y and x. This specificity is essential once the researcher wishes to make causal statements about estimated relationships or use them to perform counterfactuals. Suppose for example, the researcher has regressed winning bids on the number of bidders and estimates the coefficient on the number of bidders is \$100. Can this estimate be interpreted as the causal effect of adding another bidder to a future auction? We would argue that without further knowledge about the institutional and economic features of the auctions under study the answer is no. What separates structural models from nonstructural models, and some structural models from others, is how clearly the connections are made between institutional, economic, and statistical assumptions and the estimated relationships. While it is possible to assert that assumptions exist that make the estimated relationship causal, the plausibility of such claims ultimately rests on whether these assumptions are reasonable for the researcher's application.

As we discuss later in Section 8, IO economists have developed a variety of structural models of auction bid data. These models have been used to derive causal models of the equilibrium relations between winning bids and the number of bidders. Paarsch (1992), for example, was the first to compare empirical models of winning bids in private value and common value sealed-bid auctions. For instance, he showed that for sealed-bid auctions with risk-neutral, expected profit-maximizing, Pareto-distributed-private-value bidders would have the following density of winning bids y given a known number of

bidders, x:

$$f(y \mid x, \theta) = \frac{\theta_2 x}{y^{\theta_2 x+1}} \left[\frac{\theta_1 \theta_2 (x-1)}{\theta_2 (x-1) - 1} \right]^{\theta_2 x}.$$

Using this density, Paarsch derives the expected value of the winning bid conditional on the number of bidders:

$$E(y \mid x) = \left[\frac{\theta_1 \theta_2 (x-1)}{\theta_2 (x-1) - 1} \right] \frac{\theta_2 x}{\theta_2 x - 1}.$$

Paarsch's paper motivated IO economists to think about what types of institutional, economic and statistical assumptions were necessary to recover causal relationships from auction data. For example, researchers have asked how risk aversion, collusion and asymmetric information change the equilibrium distribution of bids. Researchers also have compared the observable implications of using sealed-bid versus open-outcry auctions.

In closing this section, we should re-emphasize the general goal of structural econometric modeling. Structural econometric models use economic and statistical assumptions to identify economic quantities from the joint density of economic data, $f(x, y)$. The main strength of this approach is that, done right, it can make clear what economic assumptions are required to draw causal economic inferences from the distribution of economic data.

3. Constructing structural models

Having introduced the concept of a structural model, we now explore how structural modelers go about constructing econometric models.

3.1. Sources of structure

There are two general sources of "structure" in structural models – economics and statistics. Economics allows the researcher to infer how economic behavior and institutions affect relationships between a set of economic conditions x and outcomes y. Often these economic models are deterministic, and as such do not speak directly to the distribution of noisy economic data. Structural econometric modelers thus must add statistical structure in order to rationalize why economic theory does not perfectly explain data. As we show later, this second source of structure may affect which economic quantities a researcher can recover and which estimation methods are preferable.

In any structural modeling effort a critical issue will be: How did the structural modeler know what choices to make when introducing economic and statistical assumptions? Most answers to this question fall into one of three categories: those made to reflect economic realities; those made to rationalize what is observed in the data or describe how the data were generated; and, those made to simplify estimation. We should

note at the outset that there is no necessary agreement among structural modelers as to how to make these choices. Some purists, for example, believe that structural models must come from fully-specified stochastic economic models. Others find it acceptable to add structure if that structure facilitates estimation or allows the researcher to recover economically meaningful parameters. For instance, economic theory may make predictions about the conditional density of y given x, $f(y \mid x)$, but may be silent about the marginal density of x, $f(x)$. In this case, a researcher might assume that the marginal density of x does not contain parameters that appear in the conditional density. Of course, there is nothing to guarantee that assumptions made to facilitate estimation are reasonable.

The "structure" in a structural model is there because the researcher explicitly or implicitly chose to put it there. Although we have argued that one of the advantages of a structural econometric model is that researchers can examine the sensitivity of the structural model estimates to alternative assumptions, this is sometimes easier said than done.

The following example illustrates how some of these issues can arise even in a familiar linear regression setting. Specifically, we ask what types of assumptions are required to interpret a regression of outputs on inputs as a production function.

EXAMPLE 1. Imagine an economist with cross-section, firm-level data on output, Q_t, labor inputs, L_t, and capital inputs, K_t. To describe the relationship between firm i's output and inputs, the economist might estimate the regression:

$$\ln Q_t = \theta_0 + \theta_1 \ln L_t + \theta_2 \ln K_t + \epsilon_t, \tag{1}$$

by ordinary least squares (OLS). In this regression, the θ's are unknown coefficients and the ϵ_t is an error term that accounts for the fact that the right-hand side input variables do not perfectly predict log output.

What do we learn by estimating this regression? Absent more information we have estimated a descriptive regression. More precisely, we have estimated the parameters of the best linear predictor of $y_t = \ln(Q_t)$ given $x_t = (1, \ln(L_t), \ln(K_t))'$ for our sample of data. Goldberger (1991, Chapter 5) provides an excellent discussion of best linear predictors. The best linear predictor of y given a univariate x is $\text{BLP}(y \mid x) = a + bx$, where $a = E(y) - bE(x)$ and $b = \text{Cov}(y, x)/\text{Var}(x)$. Absent more structure, the coefficients a and b are simply functions of population moments of $f(x, y)$.

If we add to our descriptive model the assumption that the sample second moments converge to their population counterparts

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} x_t x_t' = M_{xx} \quad \text{and} \quad \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} x_t y_t = M_{xy},$$

and that M_{xx} is a matrix of full rank, then OLS will deliver consistent estimates of the parameters of the best linear predictor function. Thus, if we are interested in predicting the logarithm of output, we do not need to impose any economic structure and very little

statistical structure to estimate consistently the linear function of the logarithm of labor and logarithm of capital that best predicts (in a minimum-mean-squared-error sense) the logarithm of output.

Many economists, however, see regression (1) as more than a descriptive regression. They base their reasoning on the observation that (1) essentially looks like a logarithmic restatement of a Cobb–Douglas production function: $Q_t = A\, L_t^{\alpha} K_t^{\beta} \exp(\epsilon_t)$. Because of the close resemblance, they might interpret (1) as a production function.

A critical missing step in this logic is that a Cobb–Douglas production function typically is a deterministic relationship for the producer, whereas the regression model (1) includes an error term. Where did the error term in the empirical model come from? The answer to this question is critical because it affects whether OLS will deliver consistent estimates of the parameters of the Cobb–Douglas production function, as opposed to consistent estimates of the parameters of the best linear predictor of the logarithm of output given the logarithms of the two inputs. In other words, it is the combination of an economic assumption (production is truly Cobb–Douglas) and statistical assumptions (ϵ_t satisfies certain moment conditions) that distinguishes a structural model from a descriptive one.

Deterministic production function models provide no guidance about the properties of the disturbance in (1). The researcher thus is left to sort out what properties are appropriate from the details of the application. One could imagine, for instance, the modeler declaring that the error is an independently-distributed, mean-zero measurement error in output, and that these errors are distributed independently of the firms' input choices. In this case, OLS has the potential to deliver consistent estimates of the production function parameters.

But how did the modeler know that ϵ_t was all measurement error? As we discuss later this is likely too strong an assumption. A more plausible assumption is that the error also includes an unobservable (to the econometrician) difference in each firm's productivity (e.g., an unobservable component of A_t in the Cobb–Douglas function). The existence of such components raises the possibility that the input choices are correlated with ϵ_t. Such correlations invalidate the use of OLS to recover consistent estimates of the parameters of the Cobb–Douglas production function.

Even if one were willing to assume that ϵ_t is measurement error distributed independently of x_t, additional economic structure is necessary to interpret the OLS parameter estimates as coefficients of a Cobb–Douglas production function. By definition, a production function gives the maximum technologically feasible amount of output that can be produced from a vector of inputs. Consequently, under this stochastic structure, unless the researcher is also willing to assert that the firms in the sample are producing along their Cobb–Douglas production function, OLS applied to (1) does not yield consistent estimates of the parameters of this production function. In a theoretical realm, the assumption of technologically efficient production is relatively innocuous. However, it may not fit the institutional realities of many markets. For example, a state-owned firm may use labor in a technologically inefficient manner to maximize its political capital with unions. Regulators also may force firms to operate off their production functions.

Both of these examples underscore the point that care must exercised to ensure that the economic model fits the institutional reality of what is being studied.

This example demonstrates what assumptions are necessary for a linear regression model to have a causal economic interpretation as a production function. First, the researcher must specify an economic model of the phenomenon under consideration, including in this case the functional form for the production function. Second, she must incorporate unobservables into the economic model. This second step should receive significantly more attention than it typically does. This is because the assumptions made about the unobservables will impact the consistency of OLS parameter estimates. Sections 4 and 5 further illustrate the importance of stochastic specifications and potential pitfalls.

3.2. Why add structure?

We see three general reasons for specifying and estimating a structural econometric model.

First, a structural model can be used to estimate unobserved economic or behavioral parameters that could not otherwise be inferred from nonexperimental data. Examples of structural parameters include: marginal cost; returns to scale; the price elasticity of demand; and the impact of a change in an exogenous variable on the amount demanded or on the amount supplied.

Second, structural models can be used to perform counterfactuals or policy simulations. In counterfactuals, the researcher uses the estimated structural model to predict what would happen if elements of the economic environment change. For example, suppose that we have estimated the demand for a product and the monopolist's cost function. We could then, with added assumptions, use these estimates to calculate how market prices and quantities would change if an identical second firm entered the monopoly market.

For these calculations to be convincing, the structural modeler must be able to argue that the structural model will be invariant to the contemplated change in economic environment. Thus, if we were to evaluate instead the effect of a regulator capping a monopolist's price, we would have to maintain that the monopolist's cost function would not change as a result of the regulation. That these assumptions need to be made and checked is again another illustration of the value of a structural model – it can help researchers identify what assumptions are required in order to draw inferences and make predictions about economic behavior and phenomena.

Finally, structural models can be used to compare the predictive performance of two competing theories. For example, we could compare the performance of quantity-setting versus price-setting models of competition. It is important to emphasize that these comparisons do not provide unambiguous tests of the underlying economic theories. Indeed, these comparisons are always predicated on untestable assumptions that are not part of the theory. For instance, any "test" of quantity-setting behavior versus price-setting

behavior is predicated on the maintained functional forms for demand, costs, and the unobservables. Thus, the only sense in which one can "test" the two theories is to ask whether one of these ways of combining the same economic and stochastic primitives provides a markedly better description of observed or out-of-sample data.

Because we cannot test economic models independent of functional form assumptions for a finite number of observations, it is important to recognize that structural parameter estimates may well be sensitive to these assumptions. For example, if we were trying to estimate consumer surplus, we should be aware that it might make a tremendous difference that we assumed demand was linear, as opposed to constant elasticity. While this sensitivity to functional form can be viewed as a weakness, it also can be viewed as a strength. This is again because the "structure" in structural models forces researchers to grapple directly with the economic consequences of assumptions.

The "structure" in structural models also can affect statistical inferences about economic primitives. Here we have in mind the impact that a researcher's functional form choices can have on the size and power of hypothesis tests. When, as is usually the case, economic theory does not suggest functional forms or what other variables might be relevant in an application, researchers will be forced to make what may seem to be arbitrary choices. These choices can have a critical impact on inferences about parameters. For example, if a researcher wants to fail to reject a null hypothesis, then she should specify an extremely rich functional form with plenty of variables that are not part of the economic theory. Such a strategy will likely decrease the power of the statistical test. For instance, if a researcher would like to fail to reject the integrability conditions for her demand functions, she should include as many demographic variables as possible in order to soak up across-household variation in consumption. This will tend to reduce the apparent precision of the estimated price coefficients and make it difficult to reject the null hypothesis of integrability. Conversely, if she would like to reject integrability, then she should include few, if any, demographic controls. This would increase the apparent precision in the price coefficients and increase the likelihood of rejection for two reasons: (1) she has reduced the number of irrelevant variables; and, (2) the effect of price may be exaggerated by the omission of relevant variables that are correlated with prices.

This discussion underscores the delicate position empiricists are in when they attempt to "test" a particular parameter or theory. For this reason, structural modelers should experiment with and report how sensitive their inferences are to plausible changes in functional forms, or the inclusion and exclusion of variables not closely tied to economic theory.

Finally, we should emphasize that these putative advantages do not always mean structural models should be favored over nonstructural models. Indeed, there are many interesting applications where there is little or no useful economic theory to guide empirical work. We certainly do not believe this should stop the collection or description of data. When on the other hand there is a substantial body of economic theory to guide empirical work, researchers should take advantage of it. In some cases, there may be a

large body of economic theory on a particular topic, but that theory may have few implications for data. In this case, structural modelers can make important contributions by making it clear what is needed to link theory to data. By being clear about what in the theory the empirical researcher can estimate, it becomes easier for economists to improve existing theory.

The advantages of structural models of course do not all come for free. All economic theories contain assumptions that are not easily relaxed. While theorists sometimes have the luxury of being able to explore stylized models with simplifying assumptions, structural econometric modelers have to worry that when they use stylized or simplifying assumptions they will be dismissed as arbitrary, or worse: insensitive to the way the world "really works". This problem is compounded by the fact that economic data rarely come from controlled experimental settings. This means that structural econometric modelers often must recognize nonstandard ways in which nonexperimental data are generated and collected (e.g., aggregation and censoring). Such complications likely will force the structural modeler to simplify. The danger in all of these cases is that the structural model can then be seen as "too naive" to inform a sophisticated body of theory. We expect that readers can see this already in Example 1.

3.3. Evaluating structure – single equation models

The standard multiple linear regression model is a useful place to begin understanding issues that arise in evaluating and interpreting structural and nonstructural models. Consider the linear regression model, $y = \alpha + x\beta + \epsilon$. The mathematical structure of this model lends an aura of "structure" as to how y is related to x. What motivates this structure? A satisfactory answer to this question minimally must address why we are regressing y on x. From a statistical perspective, we can always regress y on x or x on y. The coefficients in these regressions can then be given statistical interpretations as the coefficients of best linear predictor functions. Issues of economic causality, however, are not resolved simply because a researcher puts y on the left-hand side and x on the right-hand side of a linear equation.

Economists estimating linear regression models usually invoke economic arguments to make a case that x causes y. Assuming that the researcher has made a convincing case, what should be made of the regression of y on x? Absent an economic model showing that y and x are linearly related, all one can say is that under certain conditions ordinary least squares regression will provide consistent estimates of a best linear predictor function. The regression does not necessarily deliver an estimate of how much the conditional mean of y changes with a one unit change in x, and certainly not the causal impact of a one-unit change in x on y.

Despite this, some researchers use regression coefficient signs to corroborate an economic model in the belief that multiple regressions "hold constant" other variables. For example, a researcher might develop a deterministic economic model that shows: "when x increases, y increases". This result then becomes the economic justification for using a regression of y on x and other variables to "test" the theory. One problem

with this approach is that unless the economic model delivers a linear conditional mean specification for y given x, the regression evidence about the sign of x need not match deterministic comparative statics predictions. In general, the empirical researcher must first use economics and statistics to demonstrate that the relevant economic quantity or comparative static effect can be identified using the available data and estimation technique. To see this point more clearly, consider the following example.

EXAMPLE 2. A microeconomist has cross-section data on a large number of comparable firms. The data consist of outputs, Q, in physical units, total costs, TC, and the firms' two input prices, p_K and p_L. The researcher's goal is to learn about the firms' (by assumption) common technology of production. The researcher decides to do this by estimating one of the following regression models:

$$\begin{aligned} &\text{Model 1:} \quad \ln \mathrm{TC}_i = \theta_0 + \theta_1 \ln Q_i + \theta_2 \ln p_{Ki} + \theta_3 \ln p_{Li} + \eta_i, \\ &\text{Model 2:} \quad \ln Q_i = \beta_0 + \beta_1 \ln \mathrm{TC}_i + \beta_2 \ln p_{Ki} + \beta_3 \ln p_{Li} + \epsilon_i. \end{aligned} \tag{2}$$

These specifications differ according to whether the natural logarithm of output or the natural logarithm of total costs is a dependent or independent variable.

Which specification makes better economic sense? In an informal poll of colleagues, we found most thought Model 1 was more sensible than Model 2. The logic most often given for preferring Model 1 is that it looks like a cost function regression. When asked how to interpret the parameters of this regression specification, most say that θ_1 is an estimate of the elasticity of total cost with respect to output. As such, it provides a measure of scale economies. Those who prefer the second equation seem to base their preference on an argument that total cost is more likely to be "exogenous". To them this means that OLS is more likely to deliver consistent estimates of production or cost parameters.

How might we go about deciding which specification is correct? A structural modeler answers this question by answering two prior questions: What economic and statistical assumptions justify each model? And, do these assumptions make sense for the application at hand? In Example 5 of Section 4, we show that Models 1 and 2 can be derived from competing plausible economic and stochastic assumptions. That is, under one set of economic and stochastic modeling assumptions, we can derive Model 1. Under another set of assumptions, we can do the same for Model 2. Without knowing the details of the firms and markets being studied, it is impossible to decide which set of assumptions is more appropriate.

How do researchers only interested in data description decide which specification is correct? They too must answer prior questions. But these questions only pertain to the goals of their statistical analysis. If, for example, their goal is prediction, then they would choose between Models 1 and 2 based on the variable they are trying to predict. They then would have to decide which right-hand side variables to use and how these variables would enter the prediction equation. Here, researchers have to worry

that if their goal is post-sample prediction, they may over-fit within sample by including too many variables. While statistical model selection criteria can help systematize the process of selecting variables, it is not always clear what one should make of the resulting model.

In some cases, researchers do not have a clear economic model or descriptive criterion in mind when they estimate a regression model such as Model 1 by ordinary least squares. In this case, what can be made of the coefficient estimates obtained from regressing y on the vector x? As discussed above, ordinary least squares delivers consistent estimates of the coefficients in the best linear predictor of y given x. But what information does the BLP$(y \mid x)$ provide about the joint distribution of y and x? In general, the BLP will differ from the more informative conditional expectation of y given x, $E(y \mid x)$, which is obtained from $f(x, y)$ as $\int yf(y \mid x)\,\mathrm{d}y$. Thus, $\theta_1 = \partial\,\mathrm{BLP}(y \mid x)/\partial x_1$ in Model 1 will not in general equal how much expected log total costs will increase if we increase log output by one unit (i.e., $\partial E(y \mid x)/\partial x_1$). Only under certain conditions on the joint density of y and x are the BLP function and the conditional expectation function the same. Despite this well-known general lack of equivalence between the BLP$(y \mid x)$ and $E(y \mid x)$, many descriptive studies treat linear regression slope coefficient estimates as if they were estimates of the derivative of $E(y \mid x)$ with respect to x. Occasionally, some studies adopt the position that while the best linear predictor differs from the conditional expectation, the signs of the coefficients of the BLP$(y \mid x)$ will be the same as those of $\partial E(y \mid x)/\partial x$ provided the signs of $\partial E(y \mid x)/\partial x$ do not change with x. Unfortunately, as White (1980) demonstrates, there is no reason to expect that this will be true in general.

When the conditional expectation of y is nonlinear in x, statistical theory tells us (under certain sampling assumptions) that a regression provides a best (minimum expected squared prediction error) linear approximation to the nonlinear conditional expectation function. It is perhaps this result that some place faith in when they attempt to use regressions to validate an economic comparative static result. However, absent knowledge from economics or statistics about the joint distribution of y and x, this approximation result may be of limited value. We do not, for example, know how good the linear approximation is. We do not know if x causes y or y causes x. In sum, BLP$(y \mid x)$ and $E(y \mid x)$ are simply descriptive statistical quantities.

By making economic and statistical assumptions, however, we can potentially learn something from the linear approximation. For example, if we had an economic theory that suggested that there was a negative causal relationship between y and z, then the bivariate regression slope coefficient's sign might tell us whether the evidence is consistent with the theory. But this may be a weak confirmation of the theory and it certainly does not provide us with a sense of the strength of the relationship if the conditional mean function, $E(y \mid z)$, is nonlinear in z.

Descriptive researchers (and structural modelers) also have to worry about whether they have collected all of the data needed to examine a particular prediction about a conditional mean. Consider, for example, the case where an economic theory delivers a prediction about the conditional mean of y given x_1 and x_2, $E(y \mid x_1, x_2)$, where

y, x_1 and x_2 are scalars. Suppose that y is a customer's demand for electricity during the day, x_1 is the price of electricity during the day, and x_2 is average temperature during the day. Economic theory predicts that electricity demand is decreasing in the daily price after controlling for the average daily temperature. However, if we do not include x_2 on the right-hand side when we regress y on x_1, then we obtain the best linear approximation to $E(y \mid x_1)$, not $E(y \mid x_1, x_2)$. The difference may be very important. For instance, the function $g(x_1) = E(y \mid x_1)$ may not depend on x_1, whereas the function $h(x_1, x_2) = E(y \mid x_1, x_2)$ may depend on both x_1 and x_2.

In the usual textbook analysis of omitted variables in a linear model, it is straightforward to establish when an omitted variable will cause bias and produce inconsistent estimates. When the conditional mean is nonlinear, and we proceed as if it is linear, the familiar reasoning is not as straightforward. In addition to the omitted variable, we have to worry that even if we had included the omitted variable, that $\partial E(y \mid x_1, x_2)/\partial x_1 \neq \partial \,\mathrm{BLP}(y \mid x_1, x_2)/\partial x_1$. Absent a theory that says that y is linearly related to x_1 and x_2, the effect of omitting a relevant regressor is much harder to evaluate. Specifically, suppose

$$y_t = g(x_{1t}, x_{2t}) + \epsilon_t = g(x_t) + \epsilon_t$$

and $E(\epsilon_t \mid x_t) = 0$ so that $E(y \mid x) = g(x_t)$ is a nonlinear function. Running an ordinary least squares regression of y_t on z_t, where z_t is a vector of known functions of x_{1t} and x_{2t}, yields a consistent estimate of β where β is defined as follows:

$$y_t = z_t\beta + \big[g(x_t) - z_t\beta\big] + \epsilon_t = z_t\beta + \eta_t.$$

The parameter β is the linear combination of the z_t's that best predicts the y_t's for the population. By construction $E(z_t\eta_t) = 0$, but the partial derivative of $z_t\beta$ with respect to x_1 could differ in both sign and magnitude from the partial derivative of the conditional mean, $\partial g(x_t)/\partial x_{1t}$ depending on how well $z_t\beta$ approximates $g(x_t)$.

3.4. Evaluating structure – simultaneous equation models

The original Cowles Commission econometricians paid particular attention to developing econometric models that could represent the concept of an "economic equilibrium". Indeed, the term "structural model" often is associated with econometric models that have multiple simultaneous equations, each of which describes economic behavior or is an identity. The term simultaneous emphasizes that the left-hand side variables also can appear as right-hand side variables in other equations. The term "reduced form" was introduced to describe an alternative representation of a simultaneous system – one in which the dependent variables were explicitly represented only as functions of the x's and unobservables.

To understand what is "structural" in simultaneous equations models, it is useful to begin with a standard linear supply and demand model.

EXAMPLE 3. In a standard linear demand and supply model, the demand curve gives the quantity that consumers would like to purchase at a given price, conditional on other variables that affect demand, and the supply curve gives how much firms are willing to sell at a given price, conditional on other supply shifters. Mathematically,

$$\begin{aligned} q_t^s &= \beta_{10} + \gamma_{12} p_t + \beta_{11} x_{1t} + \epsilon_{1t}, \\ p_t &= \beta_{20} + \gamma_{22} q_t^d + \beta_{22} x_{2t} + \epsilon_{2t}, \\ q_t^s &= q_t^d, \end{aligned} \tag{3}$$

or in matrix notation:

$$[\, q_t \quad p_t \,] \begin{bmatrix} 1 & -\gamma_{22} \\ -\gamma_{12} & 1 \end{bmatrix} - [\, 1 \quad x_{1t} \quad x_{2t} \,] \begin{bmatrix} \beta_{10} & \beta_{20} \\ \beta_{11} & 0 \\ 0 & \beta_{22} \end{bmatrix} = [\, \epsilon_{1t} \quad \epsilon_{2t} \,], \tag{4}$$

$$y_t' \Gamma - x_t' B = \epsilon_t', \tag{5}$$

where Γ and B are matrices containing the unknown parameters that characterize the behavior of consumers and producers, q_t is equilibrium quantity at time t, p_t is equilibrium price, y_t is a two-dimensional vector, ϵ_t is a two-dimensional vector of unobserved random variables, and the exogenous variables, x_t, consist of a constant term, a supply shifter x_{1t} (e.g., an input price) and a demand shifter x_{2t} (e.g., household income).

To find out what restrictions the system (3) imposes on the conditional distribution of y given x, we can first solve for the endogenous variables as a function of exogenous variables and error terms. Post-multiplying both sides of (5) by Γ^{-1}, and rearranging, gives the reduced form

$$y_t' = x_t' \Pi + v_t'. \tag{6}$$

The reduced form (6) shows that equilibrium prices and quantities are linear functions of both demand and cost shifters and both demand and cost errors.

From this perspective, the individual reduced forms for equilibrium price and quantity parallel the nonstructural descriptive linear regressions discussed in the previous subsection. Some researchers, however, over-extend this logic to claim that any regression of an endogenous variable on exogenous variables is a "reduced form" regression. Thus, they would for example label a regression of price on the supply shifters x_{2t} a "reduced form".

The critical issue again in any regression of a y on x is what do we make of the estimated coefficients? Returning to the reduced form system (6), to arrive at this representation we had to first assume that the structural equations characterizing aggregate demand and supply were linear. If we did not know they were linear, then we would not know that the reduced form (6) was linear. In short, the functional form of the structural model determines the functional form of the reduced form relationship between y_t and x_t.

The economic assumption that supply equals demand also is critical to the interpretation of Π. If, for example, price floors or ceilings prevented demand from equaling

supply, then we would not obtain a standard linear model even though the underlying demand and supply schedules were linear [see Quandt (1988)].

Even when we are assured that the reduced form has the linear form (6), we cannot interpret the Π and proceed to estimation without completing the specification of the demand and supply equations. To complete the structural model, for example, the researcher could specify the joint distribution of x and y, or alternatively, as is common in the literature, the conditional distribution of y given x. Still another approach is to sacrifice estimation efficiency by imposing less structure on the joint distribution. For example, estimation could proceed assuming the conditional moment restrictions

$$E(\epsilon_t \mid x_t) = 0 \quad \text{and} \quad E\left(\epsilon_t \epsilon_t' \mid x_t\right) = \Sigma. \tag{7}$$

From these conditional moment restrictions on ϵ_t, we can deduce

$$E(v_t \mid x_t) = 0, \quad \text{and} \quad E\left(v_t v_t' \mid x_t\right) = \Omega, \tag{8}$$

where

$$\Pi = B\Gamma^{-1}, \quad v_t' = \epsilon_t' \Gamma^{-1}, \quad \text{and} \quad \Omega = \left(\Gamma^{-1}\right)' \Sigma \Gamma^{-1}. \tag{9}$$

From (9), we see that Π and the variance–covariance matrix of the reduced form errors, Ω, provide information about the structural parameters in Γ. Without restrictions on the elements of Γ, B, and Σ, however, the only restriction on the conditional distribution of y_t given x_t implied by the linear simultaneous equation model is that the conditional mean of y_t is linear in x_t and the conditional covariance matrix of y_t is constant across observations.

To summarize, a *reduced form* model exists only to the extent that the researcher has derived it from a structural economic model. If the researcher is unwilling to assume functional forms for the supply and demand equations, then the conditional means of q_t and p_t will likely be nonlinear functions of x_t, the vector of the demand and supply shifters. In this case, although we can still perform linear regressions of q_t and p_t on x_t, these linear regressions are not reduced forms. Instead, these regressions will deliver consistent estimates of the parameters of the best linear predictors of the dependent variables given x_t. How these parameter estimates are related to the price elasticity of demand or supply or other causal effects is unknown. Additionally, as discussed earlier, unless the researcher is willing to place restrictions on the functional forms of the conditional means of q_t and p_t given x_t, it will be difficult to make even qualitative statements about the properties of $E(p_t \mid x_t)$ or $E(q_t \mid x_t)$.

Notice, it is economic theory that allows us to go beyond descriptive or statistical interpretations of linear regressions. If we assume stochastic linear supply and demand equations generate y_t, and impose the market-clearing conditions $q_t^s = q_t^d$, then the equations in (9) allow us *in principle* to recover estimates of economic parameters from Π and Ω. We emphasize *in principle* because unless the values of B, Γ, and Σ can be uniquely recovered from Π and Ω, the structural model (3) has limited empirical content. Although the structural parameters are not identified, the linearity of the structural

model implies that the conditional mean of y_t is linear in x_t and the conditional variance is constant.

One might wonder how we know that the structural model given in Equation (3) is generating the observed y_t. The answer is by now familiar: Only because economic theory tells us so! Economic theory tells us what elements of x_t belong in just the supply and just the demand equations. The same theory also resolves the problem of how to identify Γ, B, and Σ from the reduced form parameters Π and Ω. Absent restrictions from economic theory, there are many different simultaneous equations models that can give rise to the same reduced form parameters Π and Ω. These models may contain radically different restrictions on the structural coefficients and impose radically different restrictions on the behavior of economic agents, yet no amount of data will allow us to distinguish among them. For economic theory to be useful, it minimally must deliver enough restrictions on Γ, B, and Σ so that the empiricist can uniquely recover the remaining unrestricted elements of Γ, B, and Σ from estimates of Π and Ω. Thus, any defense of the researcher's identification restrictions can be seen as a defense of the researcher's economic theory. Without a clearly argued and convincing economic theory to justify the restrictions imposed, there is little reason to attempt a structural econometric model.

It is well known to economic theorists that without assumptions it is impossible to derive predictions about economic behavior. For example, consumers may have preference functions and producers access to technologies. However, unless we are willing to assume, for example, that consumers maximize utility subject to budget constraints and producers maximize profits subject to technological constraints, it is impossible to derive any results about how firms and consumers might respond to changes in the underlying economic environment. An empirical researcher faces this same limitation: without assumptions, it is impossible to derive empirical results. From a purely descriptive perspective, unless a researcher is willing to assume that the joint density of x and y satisfies certain conditions, he cannot consistently estimate underlying descriptive magnitudes, such as the $\text{BLP}(y \mid x)$ or the conditional density of y given x. Further, unless an empirical researcher is willing to make assumptions about the underlying economic environment and the form and distribution of unobservables, he cannot estimate economically meaningful magnitudes from the resulting econometric model. So it is only the combination of economic and statistical assumptions that allow conclusions about economic magnitudes to be drawn from the results of an econometric modeling exercise.

Econometrics texts are fond of emphasizing the importance of exclusion restrictions for identification – an exogenous variable excluded from the equation of interest. We would like to emphasize that identification also requires inclusion restrictions – this exogenous variable must also be included in at least one equation of the structural model.

This distinction is particularly important because applied researchers are typically unwilling or unable to specify all the equations in their simultaneous equations system. This incompleteness in the econometric model reflects an incompleteness in the eco-

nomic model. This incompleteness can and should raise doubts about the validity of instruments. To see why, suppose economic theory delivers the following linear simultaneous equations model

$$y_1 = \beta y_2 + x_1\gamma + \epsilon_1,$$
$$y_2 = x_1\pi_{21} + \epsilon_2, \tag{10}$$

where the ϵ's are independently and identically distributed (i.i.d.) contemporaneously correlated errors and x_1 is a variable that is uncorrelated with ϵ_1 and ϵ_2. Suppose that a researcher is interested in estimating the structural parameters β and γ in the first equation. As it stands, these parameters are not identified. The problem is that we are missing an instrument for y_2.

What to do? One approach is to revisit the economic theory in an effort to understand where additional instruments might come from. An alternative approach that is all too common is the recommendation: "Look for an exogenous variable that is uncorrelated with the ϵ's but at the same time correlated with the right-hand side endogenous variable y_2". While these two approaches are not necessarily incompatible, the second approach does not seem to involve any economics. (This should sound a warning bell!) All one needs to find is a variable that meets a statistical criterion. In some instances, researchers can do this by searching their data sets for variables that might reasonably be viewed as satisfying this criterion.

The following suggests how a researcher might run into problems using the statistical approach: "Look for an exogenous variable that is uncorrelated with the ϵ's but at the same time correlated with the right-hand side endogenous variable y_2". Consider first the extreme case where we decide to create a computer-generated instrument for y_2 that satisfies this criterion. That is, imagine we construct an instrumental variable, x_2, as the sum of x_1 plus a computer-generated independent identically distributed random error. This new variable satisfies the statistical criteria to be a valid instrument: it is uncorrelated with the structural errors and yet correlated with y_2. Thus, it would appear that we can always identify the coefficients in the first equation as long as we have at least one exogenous variable and a good random number generator.

What is amiss here is that identification also hinges on showing that x_2 belongs in the second equation. A statistical test cannot unambiguously resolve this question (especially when x_1 and x_2 are highly correlated). However, both economics and common sense tell us that x_2 does not belong in the reduced form. Put another way, they tell us that x_2 does not belong in the structural or reduced form model – the population value of π_{22}, the coefficient associated with x_2 in the reduced form, is *zero*! Nevertheless, in finite samples we could conclude π_{22} is nonzero (and perhaps statistically so).

To understand formally why this estimation strategy fails to produce consistent estimates of β and γ, consider the instrumental variables estimator for these two parameters. This estimator uses the instruments $(x_1, x_2)'$:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T} y_{2t}x_{1t} & \sum_{t=1}^{T} x_{1t}^2 \\ \sum_{t=1}^{T} y_{2t}x_{2t} & \sum_{t=1}^{T} x_{1t}x_{2t} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{T} x_{1t}y_{1t} \\ \sum_{t=1}^{T} x_{2t}y_{1t} \end{bmatrix}.$$

A necessary condition for the consistency of this instrumental variables estimator is that the matrix

$$\frac{1}{T}\begin{bmatrix} \sum_{t=1}^{T} y_{2t}x_{1t} & \sum_{t=1}^{T} x_{1t}^2 \\ \sum_{t=1}^{T} y_{2t}x_{1t} & \sum_{t=1}^{T} x_{1t}x_{2t} \end{bmatrix}$$

converges in probability to a finite nonsingular matrix. Assume that

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} x_{1t}^2 = M_2.$$

Because $x_{2t} = x_{1t} + \eta_t$ and η_t is distributed independently of ϵ_{1t}, ϵ_{2t}, and x_{1t}, the probability limit of this matrix is equal to

$$\begin{bmatrix} M_2\pi_{21} & M_2 \\ M_2\pi_{21} & M_2 \end{bmatrix}, \tag{11}$$

which is a singular matrix. This result follows from substituting $x_{1t} + \eta_t$ for x_{2t} and $x_{1t}\pi_{21} + \epsilon_{2t}$ for y_{2t} and then applying the appropriate laws of large numbers to each element of the matrix. The singularity of (11) is just another way of saying that the rank condition for identification of the first equation of the structural model fails.

At first, this example may seem extreme. No economist would use a random number generator to create instruments – but this is our point! The researcher is informed not to do this by economics. In practice, a researcher will never know whether a specific instrument is valid. For example, our students sometimes insist that more clever choices for instruments would work. After some thought, many suggest that setting $x_2 = x_1^2$ would work. Their logic is that if x_1 is independent of the errors, so must x_1^2. Following the derivations above, and assuming that $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} x_{1t}^3 = M_3$, a finite, positive constant, we again obtain a singular matrix similar to (11), implying that this x_2 is also an invalid instrument for the same reason – it does not enter into the reduced form.

The value of economic theory is that it provides a defense for why the reduced form coefficient on a prospective instrument is not zero, i.e., the instrument is included in at least one equation of the structural model. The statistical advice that led to computer-generated instruments and x_1^2 does not do this.[2]

Some might argue that our example above ignores the fact that in most economic applications, one can find exogenous economic variables that satisfy our statistical criterion. The argument then goes on to claim that because these variables are economically related, we do not need a complete simultaneous equations model. The following example discusses this possibility.

[2] An element of x_t is a valid instrument in linear simultaneous equations model if it satisfies the conditional moment restrictions (7), $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} x_t x_t' = Q$, where Q is a positive definite matrix, and it enters at least one of the equations of the structural model. Our computer generated instrument fails this last requirement.

EXAMPLE 4. Consider a researcher who has data on the prices firms charge in different geographic markets, p_i, the number of potential demanders (population) in market i, POP_i, and whether or not the firm faces competition, COMP_i. The researcher seeks to measure the "effect" of competition on prices by regressing price on market size, as measured by the number of potential demanders and the competition dummy. That is, he estimates the regression

$$p_i = \text{POP}_i\,\theta_1 + \text{COMP}_i\,\theta_2 + \epsilon_i. \tag{12}$$

Without an underlying economic model, the OLS estimate of θ_2 on COMP_i provides an estimate of the coefficient in the best linear predictor of how prices change with the presence of competition.

The researcher might, however, claim that Equation (12) has a structural economic interpretation – namely that θ_2 measures by how much prices would change if we could introduce competition. One problem with this interpretation is that it is unlikely that the presence of competition is determined independently of price. (See Section 10.) In most entry models, competitors' decisions to enter a market are simultaneously determined with prices and quantities. In such cases, if the researcher does not observe critical demand or supply variables, then OLS will deliver inconsistent estimates of θ_2.

One possible solution to this problem is to find an instrumental variable for the presence of competitors. Suppose that the researcher claims that the average income of residents in the market, Y_i, is such an instrument. This claim might be justified by statements to the effect that the instrument is clearly correlated with the presence of competitors, as an increase in average income, holding population fixed, will increase demand. The researcher also might assert that average income is determined independently of demand for the good and thus will be uncorrelated with the error ϵ_i in Equation (12).

Does this make average income a valid instrument? Our answer is that the researcher has yet to make a case. All the researcher has done is provide a statistical rationale for the use of Y_i as an instrument exactly analogous to the argument used to justify the computer-generated instrument in Example 3. To be completely convincing, the researcher must do two more things. First, the researcher has to explain why it makes sense to *exclude* average income from Equation (12). To do this, the researcher will have to provide a more complete economic justification for Equation (12). What type of equilibrium relationship does Equation (12) represent? Why is the demand variable POP_i in this equation, but not average income, which also might be considered a demand variable? Second, the researcher also will have to make a case that Y_i enters the reduced form for COMP_i with a nonzero coefficient, or else the rank condition for identification will fail by the logic presented in Example 3. This means to be a valid instrument it must enter some other equation of the structural model. The researcher will have to be clearer about the form of the complete system of equations determining prices and the presence of competitors. This will also require the researcher to spell out the economic model underlying the simultaneous system of equations.

This next example reiterates our point that the results of a structural modeling exercise are only as credible as the economic theory underlying it. One can always impose inclusion and exclusion restrictions, but the resulting simultaneous equations model need not have a clear interpretation.

EXAMPLE 5. The 1960s' and 1970s' IO literature contains many studies that regressed firm or industry profit rates ("performance") on market concentration measures ("market structure"). In the late 1960s and early 1970s, many IO economists observed that while concentration could increase profits, there could be the reverse causation: high (low) profits would induce entry (exit). This led some to estimate linear simultaneous equations models of the form:

$$\begin{aligned} PROFIT &= \beta_0 + \beta_1 CONC + x_1\beta_2 + \epsilon_1, \\ CONC &= \alpha_0 + \alpha_1 PROFIT + x_2\alpha_2 + \epsilon_2, \end{aligned} \tag{13}$$

where *PROFIT* measures industry profitability, *CONC* measures industry concentration, the ϵ's are errors and the α's and β's are parameters to be estimated. Particular attention was paid to estimating the effect of simultaneity bias on the signs and magnitudes of α_1 and β_1.

Debates about the merits of these models often centered on what variables should be included or excluded from each equation. What proved unsatisfactory about these debates was that there were no clear answers. Put another way, although these were called "structural" models of performance and market concentration, there was no one theoretical model that provided a specific economic interpretation of α_1 and β_1. Thus, even though instrumental variable methods might deliver consistent estimates of α_1 and β_1, it was never very clear what these estimates told us about the underlying theories.

To understand why we would not call this a structural model (even though it looks like a "structural" model in the sense of having multiple endogenous variables in a single equation), consider these questions: How do we know the first equation is a behavioral relation describing how industry profitability responds to industry concentration? And: How do we know the second equation describes the way firm profitability responds to industry concentration? The population values of β_1 and α_1, the parameters that characterize how *PROFIT* responds to *CONC* and how *CONC* responds to *PROFIT*, depend crucially on which elements of x_t are included and excluded from each equation of the structural model. Unless we have an economic theory telling us which elements of x_t do and do not belong in each behavioral relation, which equation we designate as the "profit equation" and which equation we designate as a "concentration equation" is completely arbitrary. For example, we can re-write the "profit equation" in (13) as a "concentration equation",

$$\begin{aligned} CONC &= -\frac{\beta_0}{\beta_1} + \frac{1}{\beta_1} PROFIT - x_1\frac{\beta_2}{\beta_1} - \frac{1}{\beta_1}\epsilon_1 \\ &= \theta_0 + \theta_1 PROFIT + x_1\theta_3 + \eta. \end{aligned}$$

What is the difference between this "concentration equation" and the one in (13)? Because they obviously have the same left-hand side variable, the answer is that they differ in what is included on the right-hand side. One has x_1 and the other has x_2. Only economic theory can tell us which "concentration equation" is correct. That is, only economic theory can tell us what agent's behavior, or group of agents' behaviors, is characterized by a structural equation. For example, we might try to justify the profit equation in (13) as representing the profit-maximizing behavior of all firms in the industry for each level of industry concentration and conditioning variables. It is this same theory that tells us the conditioning variables in the profit equation are the x_1's and not the x_2's. Thus, economic theory also delivers the inclusion and exclusion restrictions that allow us to interpret the equations of structural econometric models.

In his criticism of large-scale macroeconometric models, Sims (1980) referred to many of the restrictions used to identify macro models as "incredible". He observed: "the extent to which the distinctions among equations in large macromodels are normalizations, rather than truly structural distinctions, has not received much emphasis" [Sims (1980, p. 3)]. By truly structural distinctions, Sims meant exclusion and other functional form restrictions derived from economic theory. This same criticism clearly applies to structural modeling of the relationship between profits and concentration. As we describe in later sections, the lack of satisfactory answers to such questions is what led some empirical IO economists to look more closely at what economic theory had to say about firm profitability and market concentration.

3.5. The role of nonexperimental data in structural modeling

Virtually all data used in empirical economic research comes from nonexperimental settings. The use of nonexperimental data can raise significant additional modeling issues for descriptive and structural modelers. Consider a researcher who wants to describe the relationship between firms' prices and the number of competitors. Suppose that the data under consideration come from markets where firms face a price cap. The most general approach to describing this relationship would be to estimate flexibly the joint distribution of prices and competitors. Provided the price cap is binding in some markets, the researcher would obtain a density that has a spike at the price cap.

Instead of flexibly estimating the joint distribution of prices and competitors, the researcher could instead use a regression to describe the relationship. As we argued earlier, OLS will deliver consistent estimates of the best linear predictor function. Suppose that absent the cap economic theory implied that the conditional mean of prices given the number of competitors (x) was linear in x. How does the presence of the cap affect the estimation of the coefficients of the conditional mean function? The answer is that the cap truncates the joint distribution and thereby alters the conditional mean of y given x. Thus, the researcher will need to model this truncation if he is to recover a consistent estimate of the coefficients in the conditional mean function.

Although similar statistical sampling issues can arise in structural models, a structural econometric modeler would view the presence of a price cap as more than a statistical nuisance. Rather, the cap is something that needs to be accounted for in the modeling of firm behavior and the unobservables.

To illustrate how structural models can account for nonexperimental data, let us return to the demand and supply model for prices and quantities. Suppose the researcher observes price, the total quantity that consumers demand at that price, and consumer income (x_1). Suppose also that the researcher has estimated the regression

$$q_t^s = \beta_0 + \beta_1 p_t + \beta_2 x_{1t} + \epsilon_{1t}$$

by OLS. For the researcher to be able to assert that they have estimated a demand curve, as opposed to a descriptive best linear predictor, they must be able to argue that price and income are uncorrelated with the error. When is this likely the case? In principle, it would be the case if the researcher could perform experiments where they faced all consumers with a random series of prices. The same experiment also could be used to estimate a supply equation using OLS, provided the researcher observed the quantity supplied at the randomly chosen price.

The key feature of the experiment that makes it possible to estimate both the demand and supply equations by OLS is that the researcher observes both the quantity demanded and the quantity supplied at each randomly chosen price. In general, the quantity demanded *will not equal* the quantity supplied at a randomly chosen price. In other words, the third equation in the demand and supply system (3) does not hold.

How do equilibrium models of price determination compare then to experimental models? One way to view nonexperimental data is that it came from a grand experiment. Imagine that in this grander experiment, the experimentalist had collected data for a vast range of randomly selected prices, incomes and input prices. Imagine now someone else extracts from the experimentalist's data only those observations in which the experimenter's randomly chosen prices, incomes and input prices resulted in the quantity supplied equaling the quantity demanded. This nonrandom sample selection would yield a data set with significantly less information and, more importantly, nonrandom prices. Thus, even though the original data came from a random experiment, the data selection process will cause OLS to no longer deliver consistent estimates of the supply and demand parameters. On the other hand, if the researcher were to apply instrumental variable techniques appropriate for a structural simultaneous equations model that (correctly) imposed the market-clearing equation (3), they would obtain consistent estimates.

Our general point here is that structural models are valuable in nonexperimental contexts because they force the researcher to grapple directly with nonexperimental aspects of data. Consider again the demand and supply model above. How did we know it was appropriate to impose $q^s = q^d$? The answer came not from a statistical model of the nonrandomness, but from our economic perspective on the nonexperimental data – we assumed that the data came from markets where there are no price floors or ceilings. Had

there been price floors or ceilings, this would change the third equation in our econometric model. For example, with binding price ceilings, we might assume that the quantity we observe is the quantity supplied. (With a binding ceiling, quantity demanded exceeds supply, but we typically would not know by how much.) Our econometric model now would have to account for this selection of quantities. A variety of such "disequilibrium" demand and supply models exist and are reviewed in Maddala (1983).

4. A framework for structural econometric models in IO

Having described differences between descriptive and structural models, we now provide a framework for building and evaluating structural econometric models. While in principle it would seem easy for empiricists to recast an economic model as an econometric model, this has not proven true in practice. The process of combining economic and statistical models is by no means formulaic. As we have indicated earlier, the process of building a tractable econometric model that respects the institutions being modeled often involves difficult trade-offs. In the remaining sections we will use the framework to illustrate the progress of structural modeling in IO.

Structural modeling, and the elements of our framework, are not new to IO or most applied fields in economics. More than fifty years ago, Trygve Haavelmo and economists at the Cowles Foundation began combining models of individual agent behavior with stochastic specifications describing what the econometrician does not know:

> The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier. ... So far, the common procedure has been to first construct an economic theory involving exact functional relationships, then to compare this theory with some actual measurements, and finally "to judge" whether the correspondence is "good" or "bad". Tools of statistical inference have been introduced, in some degree, to support such judgment ... [Haavelmo (1944, p. iii)]

While the general principle of combining economic models with stochastic specifications has been around for some time, each field of economics has had to confront its own problems of how best to combine models with data. Often the desire to have a simple, well-defined probability model of the endogenous variables forces compromises. Early on, Hood and Koopmans described the challenge facing empirical economists as:

> In reality, unobserved random variables need to be introduced to represent "shocks" in behavior relations (i.e., the aggregate effects on economic decisions of numerous variables that are not separately observed) and "errors" of measurement. The choice of assumptions as to the distribution of these random variables is further complicated by the fact that the behavior equations in question are often aggregated over firms or individuals. The implications of this fact are insufficiently explored so far. [Hood and Koopmans (1953, p. xv)]

Following in this tradition, we describe a procedure for structural economic modeling that contains three basic steps. The first step is to formulate a well-defined economic model of the environment under consideration. The second step involves adding a sufficient number of stochastic unobservables to the economic model, so that its solution produces a joint density for all observables that has positive support on all possible realizations of these variables. The final step involves verifying the adequacy of the resulting structural econometric model as a description of the observed data.

4.1. The economic model

The first main component of a structural model is a complete specification of the equations describing economic behavior, what we call the economic model. Almost all economic models in IO have the following five components:

1. A description of the economic environment, including:
   (a) the extent of the market and its institutions;
   (b) the economic actors; and
   (c) the information available to each actor.
2. A list of primitives, including:
   (a) technologies (e.g., production sets);
   (b) preferences (e.g., utility functions); and
   (c) endowments (e.g., assets).
3. Variables exogenous to agents and the economic environment, including:
   (a) constraints on agents' behavior; and
   (b) variables outside the model that alter the behavior of economic agents.
4. The decision variables, time horizons and objective functions of agents, such as:
   (a) utility maximization by consumers and quantity demanded; and
   (b) profit maximization by firms and quantity supplied.
5. An equilibrium solution concept, such as:
   (a) Walrasian equilibrium with price-taking behavior by consumers; and
   (b) Nash equilibrium with strategic quantity or price selection by firms.

While the rigor of mathematics forces theorists to be clear about these components when they build an economic model, structural econometric models differ considerably in the extent to which they spell out these components. Our later discussions will illustrate the value of trying to make these components clear. In particular, we will focus attention on component 5, the equilibrium solution concept, because this is the most critical and specific to IO models.

4.2. The stochastic model

The next step in structural modeling is unique to empirical research. It receives much less attention than it deserves. This step is the process by which one transforms a deterministic (or stochastic) economic model into an econometric model. An econometric

model is distinct from an economic model in that it includes unobservables that account for the fact that the economic model does not perfectly fit observed data. Our main point is that the process of introducing errors should not be arbitrary. Both the source and properties of these errors can have a critical impact on the distribution of the observed endogenous variables and estimation.

The four principal ways in which a researcher can introduce stochastic components into a deterministic economic model are:

1. researcher uncertainty about the economic environment;
2. agent uncertainty about the economic environment;
3. optimization errors on the part of economic agents; and
4. measurement errors in observed variables.

This subsection emphasizes how these stochastic specifications differ, and in particular how they can affect the manner by which the researcher goes about estimating structural parameters.

4.2.1. Unobserved heterogeneity and agent uncertainty

A researcher's uncertainty about the economic environment can take a variety of forms. These different forms can have dramatically different implications for identification and estimation. For this reason it is critical for structural modelers to explain where error terms come from and whose uncertainty they represent. A critical distinction that needs to be drawn in almost every instance is: Is the uncertainty being introduced shared by the economic actors and econometrician?

A common assumption is that the researcher knows much less about the economic environment than the economic agents. In this case, the economic agents base their decisions on information that the researcher can only include in an error term. For example, if the researcher did not observe auction bidders' private information about an object, then the researcher would be forced to model how this unobservable information impacted bids. In general, we refer to a situation where agents' decisions depend on something the economist does not observe as a case of unobserved heterogeneity.

Of course researchers and economic agents can share uncertainty about the economic environment under study. For instance, the bidder may know their value for an object, but not the private values of the other bidders. In each of these cases, the firm or agent is presumed to know the distribution of uncertainty and make decisions that optimize the expected value of an objective function.

It might seem that because the econometrician is ignorant in both cases that unobserved heterogeneity and agent uncertainty are two sides of the same coin – they both rationalize introducing error terms in a structural model. The distinction, however, often is important for determining which estimation procedure is appropriate. To underscore this point, we now return to the two models described in (2). We shall show that, de-

pending on our assumptions about the source of the errors, it may be appropriate to regress ln TC on ln Q and other controls, or ln Q on ln TC and these same controls.

EXAMPLE 6. Imagine that we have cross-section data on comparable firms consisting of output, Q, total costs, TC, and input prices, p_K and p_L. Our goal is to estimate α and β in the Cobb–Douglas production function

$$Q_i = A_i L_i^{\alpha} K_i^{\beta}$$

consistently, where the subscript i denotes the ith firm. Because we do not have labor and capital information, we need to derive a relationship between total costs and output. There are many possible ways of doing this, each depending on what additional assumptions we make about the economic environment in which firms make their decisions.

Suppose, for example, that the firms are in a regulated industry, and have different A_i. For the purposes of exposition, assume that demand is completely inelastic. Consider now the case of pure unobserved heterogeneity (Type 1 shocks), where A_i is observed by the firm and the regulator, but not the econometrician. For simplicity, assume the A_i are i.i.d. positive random variables. Firm profits equal:

$$\pi(p_i, K_i, L_i) = p_i A_i L_i^{\alpha} K_i^{\beta} - p_{Ki} K_i - p_{Li} L_i.$$

Suppose that the regulator chooses p_i, the price of firm i's output first, and the firm then chooses K_i and L_i. Because demand is inelastic, a regulator interested in maximizing consumer welfare will set the firm's output price equal to the minimum average cost of producing Q_i. At this price, p_i^r, the firm chooses its inputs to minimize costs given the regulator's price and Q_i. That is, the firm maximizes

$$\pi(K_i, L_i) = p_i^r A_i L_i^{\alpha} K_i^{\beta} - p_{Ki} K_i - p_{Li} L_i.$$

Solving the firm's profit-maximizing problem, yields the total cost function:

$$\mathrm{TC}_i = C_0 p_{Ki}^{\gamma} p_{Li}^{1-\gamma} Q_i^{\delta} A_i^{-\delta}, \tag{14}$$

relating firm i's observed total cost data to its output. In this equation, $\delta = 1/(\alpha + \beta)$ and $\gamma = \beta/(\alpha + \beta)$. We can transform this total cost function into a regression equation using natural logarithms:

$$\ln \mathrm{TC}_i = \ln C_0 + \gamma \ln p_{Ki} + (1 - \gamma) \ln p_{Li} + \delta \ln Q_i - \delta \ln A_i. \tag{15}$$

While this equation holds exactly for the firm, the researcher does not observe the A_i. The researcher thus must treat the efficiency differences as unobservable in this logarithm of total cost regression:

$$\ln \mathrm{TC}_i = C_1 + \gamma \ln p_{Ki} + (1 - \gamma) \ln p_{Li} + \delta \ln Q_i - \delta \ln u_i. \tag{16}$$

This regression equation contains the mean zero error term

$$\ln u_i = \ln A_i - E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln Q_i].$$

The new constant term $C_1 = \ln C_0 + E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln Q_i]$ absorbs the nonzero conditional mean of the efficiency differences. Because the A_i are i.i.d., the conditional expectation reduces to the unconditional expectation, $E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln Q_i] = E[\ln A_i]$.

To summarize, we have derived a regression equation that is linear in functions of the (regulated) firm's production parameters. The relationship includes an error term that represents the firms' unobserved productive efficiencies. This error term explains why, at the same output level and input prices, firms could have different total costs. What is left to explain, is how a researcher would estimate the production parameters. This is a nontrivial issue in general. Here it is possible to argue that under fairly weak assumptions on the distribution of the u_i we can use ordinary least squares (OLS) to recover the production parameters. Note that OLS is appropriate because we have assumed that the regulator (and not the firm) picks price to recover the firm's minimum production cost to serve output Q_i. Put another way, OLS works because the unobserved heterogeneity in firms' production efficiencies is unrelated to the left-hand side regressors: firm output (which is inelastically demanded) and input prices (inputs are elastically supplied).

Now suppose that we observe the same data, but that the firm, like the econometrician, does not know its productive efficiency, A_i. This assumption leads to a different estimation strategy. In this case, the firm now must make its input decisions before it knows A_i. As long as the firm cannot undo this choice once A_i is realized, the firm maximizes expected profits taking into account the distribution of A_i. Now firm i's expected profit function is:

$$E\big[\pi(p_i, L_i, K_i)\big] = E\big[p_i A_i L_i^{\alpha} K_i^{\beta}\big] - p_{Ki} K_i - p_{Li} L_i. \tag{17}$$

We should note here that the expectation operator represents the firm's expectation.

Assume that the regulator again chooses p_i; the firm then chooses K_i and L_i. For simplicity, suppose that the regulator and the firm have the same uncertainty about the firm's productive efficiency. Suppose additionally that the regulator sets price, p_i^{er}, such that the firm earns zero profits in expectation. The firm then maximizes:

$$E\big[\pi\big(p_i^{\mathrm{er}} K_i, L_i\big)\big] = p_i^{\mathrm{er}} E\big[A_i L_i^{\alpha} K_i^{\beta}\big] - p_{Ki} K_i - p_{Li} L_i. \tag{18}$$

The first-order conditions for expected profit maximization imply

$$L_i = \left[\frac{\alpha p_{Ki}}{\beta p_{Li}}\right] K_i. \tag{19}$$

Observed total costs therefore equal

$$\mathrm{TC}_i = \frac{\alpha + \beta}{\beta} p_{Ki} K_i \tag{20}$$

and do not depend on the firm's (random) efficiency parameter A_i. Substituting these two expressions into the production function, we obtain an equation relating the observed (random) output Q_i^a to the firm's input prices and total costs

$$Q_i^a = D_0 \,\mathrm{TC}_i^{\alpha+\beta}\, p_{Ki}^{-\beta} p_{Li}^{-\alpha} A_i. \tag{21}$$

From both the firms' and the econometrician's perspective, the sole source of randomness here is the efficiency parameter A_i. Taking natural logarithms of both sides we obtain a regression equation that is linear in the production parameters

$$\ln Q_i^a = \ln D_0 + (\alpha + \beta) \ln \mathrm{TC}_i - \beta \ln p_{Ki} - \alpha \ln p_{Li} + \ln A_i. \tag{22}$$

This equation exactly explains firm i's realized production Q_i^a (which differs from the inelastically demanded quantity Q_i). Neither the firms nor the econometrician knows the A_i *ex ante*. Because the researcher also does not observe the efficiencies *ex post*, she must treat the efficiencies as random errors. She thus estimates the regression

$$\ln Q_i = D_1 + (\alpha + \beta) \ln \mathrm{TC}_i - \beta \ln p_{Ki} - \alpha \ln p_{Li} + \eta_i, \tag{23}$$

where $\eta_i = \ln A_i - E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln \mathrm{TC}_i]$. The constant term $D_1 = \ln D_0 + E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln \mathrm{TC}_i]$ absorbs the nonzero conditional mean of the efficiency differences. Once again, using the i.i.d. assumption on the A_i, the conditional expectation on $\ln A_i$ simplifies to the unconditional expectation. We can now use OLS to estimate the production parameters because by assumption the uncertainty in production is realized after the firm makes its production decision and is unrelated to total costs and input prices.

This example illustrates how the structural model's economic and stochastic assumptions can have a critical bearing on the consistency of a particular estimation strategy. Under one set of economic and stochastic assumptions, OLS applied to Equation (16) yields consistent estimates of the parameters of the firm's production function; under another set, we swap the dependent variable for one independent variable. Both models assumed (expected) profit-maximizing firms and (expected) welfare-maximizing regulators. In the first case, the stochastic shock represented only the researcher's ignorance about the productivity of firms. In the second, case, it represented uncertainty on the part of the firm, the regulator and the researcher about the productivity of the firm.

This example illustrates our initial point that a researcher should decide between models based upon how well their economic and stochastic assumptions match the environment in which the researcher's data were generated. Because no economic model is perfect in practice, the researcher often will be left choosing among imperfect assumptions and models. No statistical test can tell which model is correct. In later sections, we will discuss in more detail how a researcher might go about choosing among competing models.

4.2.2. Optimization errors

The third type of error listed above, optimization error, has received the least attention from structural modelers. In part, optimization errors have received less attention because there are few formal decision-theoretic models of optimization errors. The errors we have in mind are best illustrated by the behavior of economic agents in experiments. Experimental subjects often make errors, even when faced with relatively simple tasks.

Experimentalists' interpretations of these errors has been the source of considerable debate (e.g., see Camerer's (1995) survey). Here, we adopt a narrow view of what optimization error means so that we can illustrate the potential significance of such errors for structural models.

EXAMPLE 7. This example narrowly interprets optimization errors as the failure of agents' decisions to satisfy exactly first-order necessary conditions for optimal decisions. We are silent here on what causes this failure, and focus instead on its consequences. As an example, consider the standard consumer demand problem with unobserved heterogeneity in the utility function:

$$\min_{\lambda \geqslant 0}\Big[\max_{x \geqslant 0} U(x, \eta) + \lambda(M - p'x)\Big], \tag{24}$$

where x is an n-dimensional vector of consumption goods, p is the vector of prices, and M is the consumer's total budget. The vector η represents elements of individual tastes that the researcher does not observe. The normal first-order condition for x_i, assuming η is known is

$$\frac{\partial U}{\partial x_i}(x_i, \eta_i) - \lambda_i p_i = 0. \tag{25}$$

These equations yield the $i = 1, \ldots, n$ Marshallian demands, $x_i(p, M, \eta)$. In this case, the agent's first-order conditions are assumed to hold with probability one, so that for all realizations of η, all of the integrability conditions hold for the $x_i(p, M, \eta)$.

Now suppose that we introduce an additional source of error into the agent's demands. Although there are several ways to introduce error, imagine the errors do not impact the consumer's budget constraint (i.e., we still have $M = \sum_{i=1}^{n} p_i x_i$), but do impact the first-order conditions (25). Specifically, suppose

$$\frac{\partial U}{\partial x_i}(x, \eta) - \lambda p_i \nu_i = 0. \tag{26}$$

The researcher does not observe the ν_i, and thus treats them as random variables. Suppose for convenience that the researcher believes these errors have positive support and a mean of one in the population, so that on average the first-order conditions are correct.

How do the ν_i impact agents' decisions? If we solve the first-order conditions, and use the budget constraint, we obtain the Marshallian demand functions $x_i(p, M, \eta, \nu)$. Although the "demand curves" that result from this process satisfy homogeneity of degree zero in prices and total expenditure, they do not necessarily have a negative semi-definite Slutsky matrix for all realizations of the vector ν.

The next example shows how optimization errors can be used to rationalize why two seemingly identical consumers who face the same prices may purchase different amounts of x and y.

EXAMPLE 8. Imagine that we have demand data from a cross-section of similar consumers, all of whom have the same budget M, which they spend on two goods x and y. How should we model the differences in their consumption? One possible modeling strategy would be to say consumers have different preferences. Another would be to assume consumers have the same preference function, but that they make optimization errors when they make decisions.

Suppose each consumer has the utility function $U(x, y) = x^a y^b$ and that the first-order conditions have the form given in (26). Solving the first-order conditions, yields

$$\frac{a}{x} = \lambda p_x \nu_{xi}, \qquad \frac{b}{y} = \lambda p_y \nu_{yi}, \qquad p_x x + p_y y = M, \tag{27}$$

where λ is the Lagrange multiplier associated with the budget constraint and ν_{xi} and ν_{yi} are positive random variables representing optimization errors for consumer i. Further algebra yields

$$\lambda = \frac{\alpha_i + \beta_i}{M} \quad \text{with } \alpha_i = \frac{a}{\nu_{xi}} \text{ and } \beta_i = \frac{b}{\nu_{yi}}, \tag{28}$$

$$x = \frac{\alpha_i}{\alpha_i + \beta_i} \frac{M}{p_x} \quad \text{and} \quad y = \frac{\beta_i}{\alpha_i + \beta_i} \frac{M}{p_y}. \tag{29}$$

These demand functions look exactly like what we would get if there were no optimization error, and we had instead started with the Cobb–Douglas utility function $U(x, y) = x^{\alpha_i} y^{\beta_i}$. In other words, if we had started the modeling exercise by assuming that consumers did not make optimization errors, but instead had Cobb–Douglas preferences with heterogeneous utility parameters, we would have obtained an observationally equivalent demand model. The only way we might be able to distinguish between the two views would be to have data on consumers' choices across different purchase occasions. In this case, if consumers' tastes were time invariant, but their optimization errors varied intertemporally, we could in principle distinguish between optimization error and unobserved heterogeneity in tastes.

Optimization errors also can reduce the perceived rationality of agents' behavior. The following example shows that the way in which optimization errors are introduced can affect the extent to which firms are observed to be optimizing.

EXAMPLE 9. Consider a set of firms that have the common production function $Q = L^\alpha K^\beta$. Suppose each firm makes optimization errors when it attempts to minimize production costs. Specifically, assume that the factor demand functions are generated by solving the following three equations:

$$p_L \nu_L = \lambda \alpha K^\beta L^{\alpha-1}, \quad p_K \nu_K = \lambda \beta K^{\beta-1} L^\alpha, \quad \text{and} \quad Q = K^\beta L^\alpha, \tag{30}$$

where λ is the Lagrange multiplier associated with the constraint that the firm produces using the production function, and ν_{Li} and ν_{Ki} are unit mean, positive random variables representing optimization errors for firm i. Solving these three equations yields

following two factor demands:

$$L = Q^{\frac{1}{\alpha+\beta}} \left[\frac{p_K}{p_L}\right]^{\frac{\beta}{\alpha+\beta}} \left[\frac{\beta v_K}{\alpha v_L}\right]^{\frac{\beta}{\alpha+\beta}}, \tag{31}$$

$$K = Q^{\frac{1}{\alpha+\beta}} \left[\frac{p_K}{p_L}\right]^{\frac{-\alpha}{\alpha+\beta}} \left[\frac{\beta v_K}{\alpha v_L}\right]^{\frac{-\alpha}{\alpha+\beta}}. \tag{32}$$

An implication of the optimization errors, v_{Li} and v_{Ki}, is that the symmetry restriction implied by cost-minimization behavior fails. Specifically, the restriction

$$\frac{\partial L}{\partial p_K} = \frac{\partial K}{\partial p_L} \tag{33}$$

does not hold. Consequently, despite the fact that factor demands honor the feasibility constraint implied by the production function, they do not satisfy all of the restrictions implied by optimizing behavior.

Depending on how optimization errors are introduced, varying degrees of rationality can be imposed on factor demand and consumer demand systems. For example, optimization errors can be introduced in such a way as to yield consumer demands that satisfy the budget constraint and nothing else. This is another way of making Gary Becker's (1962) point that much of the apparent rationality in economic behavior comes from imposing a budget constraint or a technological constraint on what otherwise amounts to irrational behavior.

This discussion of optimization errors has hopefully demonstrated the extremely important and often overlooked point: the addition of disturbances to deterministic behavioral relationships is not innocuous. Depending on how this is done, a well-defined deterministic economic model can be transformed into an incoherent statistical model. For example, if the random disturbances in Equation (26) are allowed to take on values less than zero, for certain realizations of v this system of first-order conditions may not have a solution in x and λ, or may have multiple solutions. Because of these concerns, we recommend that the underlying economic model be formulated with the stochastic structure included, rather than including random shocks into a deterministic model as an afterthought.

4.2.3. Measurement error

Besides these sources of error, structural models also may include measurement errors. Measurement errors occur when the variables the researcher observes are different from those the agents observe. In most cases, it is impossible for researchers to distinguish measurement error from the three other sources of error. As we shall see below, this distinction is nevertheless important, having significant implications not only for estimation and testing, but also for policy evaluations.

Measurement errors also occur in exogenous variables. Unfortunately, these measurement errors often are ignored even though they can be a much greater source of concern. For example, measurement errors in the regressors of a linear regression model will destroy the consistency of OLS. Attempts to handle measurement error in exogenous variables often are frustrated by the fact that there typically is little prior information about the properties of the measurement error. This means that the researcher must predicate any solution on untestable assumptions about the measurement error. As a result, most researchers only acknowledge measurement error in an exogenous variable when they think that the measurement error constitutes a large component of the variation in the exogenous variable.

Measurement error can serve useful purposes in structural econometric modeling. For example, measurement error can make what would otherwise be an incoherent structural model coherent. Consider the case where consumers face nonlinear budget sets. Suppose a consumer must pay $1 per unit for the first 10 units consumed and then $10 per unit for all units beyond the tenth unit consumed. Given the large difference in price between the tenth and eleventh units, we would expect that many consumers would purchase exactly 10 units. In real data, we often do not see dramatic spikes in consumption when marginal prices increase. One way to account for this is to assume that actual consumption is measured with error. This is consistent with the theoretical model's prediction of a probability mass at exactly 10 units, but our not observing a large number of consumers consuming exactly ten units.

Measurement error also is a straightforward way of converting a deterministic economic model into a statistical model. In Example 1, for instance, we introduced measurement errors to justify applying OLS to what otherwise should have been a deterministic relation. However, as we also noted in Example 1, it is usually unrealistic to assume that measurement error is the only source of error. In general, measurement error should be introduced as one of several possible sources of error.

4.3. Steps to estimation

Given a well-defined stochastic model, the next part of our framework is to add any parametric and distributional assumptions necessary to finalize the model. The researcher then is in a position to select an estimation technique and to formulate, where possible, tests of maintained assumptions. We think of this process as having four interrelated selections:

1. selection of functional forms;
2. selection of distributional assumptions;
3. selection of an estimation technique; and
4. selection of specification tests.

There are several criteria a researcher should keep in mind when choosing a functional form. One of the most important is that there is a trade-off between data availability and parametric flexibility. Larger datasets usually allow greater parametric flexibility.

A second criterion is that the functional form should be economically realistic. To take an extreme example, if we are interested in estimating an input elasticity of substitution, then a Cobb–Douglas production function will not work. While this is an extreme case, the structural modeling literature contains nontrivial examples where the functional form almost entirely delivers the desired empirical result.

A third criterion is ease of estimation. If a specific functional form results in a model that is easier to estimate, that should certainly be a factor in its favor. Similarly, if one functional form makes it easier to impose economic restrictions than another, then that too should favor its selection. As an example, it is very easy to impose homogeneity of degree one in input prices on a translog production function. This is not the case for a quadratic cost function. A final criterion is estimation transparency. In some cases, it pays to select a functional form that leads to simpler estimation techniques. This has the advantage of making it easier for other researchers to understand how the researcher arrived at their estimates.

Turning now to the choice of distributional assumptions, a researcher's stochastic specification may or may not involve a complete set of distributional assumptions. To the extent that the researcher is willing to completely specify the distribution of the model errors, the structural model implies a conditional distribution of the observed endogenous variables given the exogenous variables. At this point the researcher can consider using maximum likelihood, or a similar technique (e.g., simulated maximum likelihood or the EM algorithm) to estimate the parameters of interest.

As a specific example, consider an optimizing model of producer behavior. Suppose the economic model specifies a functional form for $\pi(y, x, \epsilon, \beta)$ – a firm's profit function as a function of outputs produced and inputs consumed, y; a vector of input and output prices, x; the vector of firm characteristics observable to the firm but not the researcher, ϵ; and a vector of parameters to be estimated, β. If the firm maximizes profits by choosing y, we have the first-order conditions

$$\frac{\partial \pi(y, x, \epsilon, \beta)}{\partial y} = 0. \tag{34}$$

Assuming that the inverse function $y = h(x, \epsilon, \beta)$ exists and assuming the only source of error, ϵ, has the density, $f(\epsilon, \theta)$, we can apply the change-of-variables formula to compute the density of y from the density of the unobservable ϵ

$$p(y \mid x, \theta, \beta) = f\big(h^{-1}(y, x, \beta), \theta\big)\left|\frac{\partial h^{-1}(y, x, \beta)}{\partial y}\right|. \tag{35}$$

This density can be used to construct the likelihood function for each observation of y.

The final two items on our list include familiar issues in estimation and testing. An advantage of using maximum likelihood in the previous example is that it would be clear to other researchers how the elements of the economic and stochastic models led to the estimation method. There are of course costs to being this complete. One is that maximum likelihood estimators may be difficult to compute. A second is that there is a trade-off between efficiency and robustness. Maximum likelihood techniques may

be inconsistent if not all of the distributional assumptions hold. Generalized method of moments and other estimation techniques may impose fewer restrictions on the distribution of ϵ, but also may yield less efficient estimates. It also is the case that alternatives to maximum likelihood may not allow the estimation of some parameters. This is a corollary to our earlier point about structure. In some instances, the researcher's economic structure exists only because of distributional assumptions. In subsequent sections, we will illustrate how distributional assumptions can identify economic primitives.

Once the researcher obtains estimates of the structural model, it is important to examine, where possible, any restrictions implied by a structural model's economic and stochastic assumptions. In addition, it is useful to examine, where possible, how sensitive estimates are to particular assumptions. Thus, if the researcher has used instrumental variable methods to estimate a model, and there are over-identifying restrictions, then these restrictions should be tested. If a researcher assumes an error term is white noise, then tests for heteroscedastic and/or autocorrelated errors are appropriate. As for the sensitivity of estimates, the researcher can check whether additional variables should be included, or whether other functional form assumptions are too restrictive. Although it is extremely difficult to determine the appropriate nominal size for these specification tests, it is still worthwhile to compute the magnitude of these test statistics to assess the extent to which the structural model estimated is inconsistent with the observed data. Once the structural model is shown not to be "wildly" inconsistent with the observed data, the researcher is ready to use this structural model to answer the sorts of questions discussed in Section 2 and this section.

4.4. Structural model epilogue

An important premise in what follows is that no structural analysis should go forward without a convincing argument that the potential insights of the structural model exceed the costs of restrictive or untestable assumptions. Knowing how to trade off these costs and benefits is critical to knowing whether it makes sense to develop and estimate a structural econometric model. We hope that our framework and our discussion of the IO literature will provide some sense of the "art" involved in building and evaluating structural models.

In what follows, we propose to show how researchers in IO have used structural econometric models. Our purpose is not to provide a complete survey of IO. There already are several excellent literature surveys of areas such as auctions and firm competition. We propose instead to provide a sense of how IO empiricists have gone about combining game-theoretic economic models and statistical models to produce structural econometric models. We also aim to provide a sense of how far IO researchers are in solving important econometric issues posed by game-theoretic models. In our discussions, we hope to convey that structural modeling should be more than high-tech statistics applied to economic data. Indeed, we aim to show through examples how the economic question being answered should motivate the choice of technique (rather than the other way around).

5. Demand and cost function estimation under imperfect competition

In this section, we discuss Porter's (1983) empirical model of competition in an oligopoly market. We begin with Porter's model for several reasons. First, it was one of the first to estimate a game-theoretic model of competition. Second, the model bears a strong resemblance to the classical demand and supply model we discussed in Section 3. Third, we think it is an excellent example of how structural econometric modeling should be undertaken. In the process of reviewing his model, we hope to illustrate how our framework can help identify the essential ingredients of a structural model.

5.1. Using price and quantity data to diagnose collusion

One of the most important research topics in IO is how to measure the extent of competition in an industry. This question is of more than academic interest, as policy makers and the courts often are called upon to assess the extent of intra-industry competition. Additionally, when policymakers or the courts find there is insufficient competition, they must go a step further and propose remedies that will prevent firms from colluding or exercising excessive unilateral market power.

Economists seek to infer the presence or absence of competition from other data, most frequently data on prices and quantities. Sometimes these studies are conducted using firm-level or product-level price and quantity information, and sometimes economists only have industry price and quantity data. The central message of the next several sections is:

> The inferences that IO researchers' draw about competition from price and quantity data rest on what the researchers assume about demand, costs, and the nature of firms' unobservable strategic interactions.

It is therefore essential to evaluate how each of these components affects a researcher's ability to use nonexperimental price and quantity data to identify the extent of competition in an industry.

The demand specification plays a critical role in competition models because its position, shape and sensitivity to competitors' actions affects a firm's ability to mark up price above cost. The IO literature typically draws a distinction between demand models for homogeneous products and differentiated products. In this section we consider homogeneous product models in which firms' products are perfect substitutes and there is a single industry price. In this case, industry demand has the general form

$$Q = h(P, Z, \beta, \nu), \tag{36}$$

where Q is total industry quantity, P is industry price, Z are market demand variables, β are parameters that affect the shape and position of market demand, and ν is a market demand error. This demand function is an economic primitive. By itself it tells us nothing about firm behavior or the extent of competition. Inferences about the extent

of competition, however, are inextricably linked to what the researcher assumes about demand. This is because the demand curve enters into firms' profit-maximizing quantity or price decisions.

To model firms' price or quantity decisions, the researcher must first take a stand on the form of firms' profit functions, specifically the functional forms of market demand and firm-level costs. Once these are specified, the researcher must then introduce assumptions about how firms interact (e.g., Cournot versus Bertrand). Combining these assumptions with the assumption of expected profit-maximizing behavior yields first-order conditions that characterize firms' optimal price or quantity decisions. This "structure" in turn affects the industry "supply" equation that the researcher would use to draw inferences about competition.

In some, but not all, cases it is possible to parameterize the impact of competition on firms' first-order conditions in such a way that they aggregate to an industry price or "supply" equation:

$$P = g(Q, W, \theta, \eta), \tag{37}$$

where W are variables that enter the firms' cost functions, θ are parameters that affect the shape and position of the firms' cost curves and possibly describe their competitive interactions, and η is an error term.

Equations (36) and (37) look like nonlinear versions of the simultaneous linear equations in (3) of Example 3. Both sets of equations describe equilibrium industry prices and quantities. The chief difference is that in an oligopolistic setting, the "supply" equation is not an aggregate marginal cost curve but an aggregation of firm first-order conditions for profit-maximization in which firms mark price up above marginal cost. The extent to which price is above marginal cost depends on firms' competitive interactions. The critical issue is: What about the demand and "supply" equations identifies the extent of competition from observations on prices and quantities?

Porter's study provides a useful vehicle for understanding the assumptions necessary to identify the extent of competition from industry price and quantity data. In particular, his study makes it clear that without imposing specific functional form restrictions on market demand and industry supply, we have no hope of estimating the market demand curve or firm cost curves. This is because the researcher only observes pairs of prices and quantities that solve (36) and (37). Even when the researcher is willing to make distributional assumptions about the joint density of ν and η, without assumptions on the functional form of (36) and (37), the assumption that P and Q are equilibrium magnitudes only implies that there is conditional density of P and Q given Z and W. Consequently, if the researcher is unwilling to make any parametric assumptions for the demand and supply equations, he would, at best, be able to only recover the joint density of P and Q given Z and W using the flexible smoothing techniques described earlier. Only by making parametric assumptions for the supply and demand equations can these two equations be separately identified and estimated from market-clearing prices and quantities. This is precisely the strategy that Porter (1983) and all subsequent

researchers take in estimating the competitiveness of a market from equilibrium price and quantity data.

Rosse (1970) first estimated the extent of unilateral market power possessed by a firm from market-clearing price and quantity, using a sample of monopoly markets. Porter's 1983 study of nineteenth century US railroad cartels is one of the first papers in IO to devise a structural econometric model of a cartelized industry.[3] The economic logic for Porter's empirical model comes from Green and Porter (1984). Green and Porter explore the idea that cartels might use price wars to discipline members who deviate from cartel prices or output quotas. Specifically, Green and Porter develop a dynamic model of a homogeneous product market in which potential cartel members face random shocks to industry demand. By assumption, firms never perfectly observe demand or other firms' output decisions. In this noisy environment, cartel participants have trouble identifying whether lower prices are the result of a breakdown in the cartel or low demand. Green and Porter's work shows that firms can support a cartel by agreeing to a period of competitive pricing of a pre-determined length whenever market prices fall below a trigger price.

In what follows, we use our framework to discuss the components of Porter's model. In particular, we focus on the assumptions that allow Porter to identify competitive pricing regimes. In the process, we hope to illustrate many of our earlier points about structural models. The main lessons we take away from Porter's analysis is that it is impossible to identify the extent of market power exercised by a firm or in an industry from a descriptive data analysis. It is also impossible to determine definitively whether firms are colluding from this sort of data analysis. Inferences about the extent of market power exercised, or the presence and pervasiveness of collusion, rest heavily on untestable economic, functional form and stochastic assumptions. In general, it is not possible to test all these assumptions. The strength of Porter's equilibrium model in which the cartel switches between monopoly and competitive prices is that it is possible to see what is needed to identify monopoly versus competitive regimes.

5.2. The economic model

5.2.1. Environment and primitives

Porter begins, as does most of the structural IO literature, by outlining a static, homogeneous product oligopoly model where the number of firms (entrants) N is exogenously given. All firms know the functional form of market demand and each others' costs. In Porter's homogeneous product model, there is a single, constant elasticity industry demand curve at each period t:

$$\ln Q_t = \alpha + \epsilon \ln P_t + Z_t'\gamma + \nu_t, \tag{38}$$

[3] See Bresnahan (1989) for a detailed survey of early work on estimating market power.

where Q is industry output, P is industry price, Z is a vector of exogenous demand shifters, γ is a conformable vector of unknown coefficients, ϵ is a time-invariant price elasticity of demand, and ν_t is an error term. It appears that Porter uses a constant elasticity demand function because it considerably simplifies subsequent calculations and estimation. Data limitations also limit Z_t to one exogenous variable, a dummy for whether competing shipping routes on the Great Lakes were free of ice. Although he does not discuss the source of the demand error term, it plausible to imagine that it is included to account for demand factors observable to firms but not to Porter.

Each firm has fixed costs of F_i and a constant elasticity variable cost function of the form

$$C_i(q_{it}) = a_i q_{it}^{\delta}, \tag{39}$$

where i indexes firms, t indexes time and q is firm-level output. The motivation for this firm-level cost function appears to be that it delivers an industry "supply" or output curve for a range models of competition.

Porter leaves portions of the economic environment unspecified. Although competing shippers are mentioned, their impact on the railroads is not explicitly modeled. Similarly, although entry by railroads occurs during the sample, the entry decisions are not modeled. (Entry is accounted for by an exogenous shift in the industry supply curve.) Finally, although Porter does not include unobservables in the individual cost functions, as we show below it is possible to rationalize part of the error term that he includes in the industry supply curve as a variable cost component common to all firms that he does not observe.

5.2.2. Behavior and optimization

Porter assumes that each period (one week), firms maximize their per-period profits choosing shipping quantities, q_{it}. Additionally, he assumes each firm forms a conjecture about how other firms will respond to changes in its quantity during that week, θ_{it}. From these behavioral assumptions, Porter derives the standard marginal revenue equals marginal cost quantity-setting first-order conditions for profit maximization by each firm:

$$p_t\left(1 + \frac{\theta_{it}}{\epsilon}\right) = a_i \delta q_{it}^{\delta-1}. \tag{40}$$

Here

$$\theta_{it} = \frac{\partial Q_t}{\partial q_{it}} \frac{q_{it}}{Q_t} = \left(1 + \frac{\partial Q_{-it}}{\partial q_{it}}\right) \frac{q_{it}}{Q_t}$$

and $Q_{-it} = \sum_{k \neq i}^{M} q_{kt}$ is the total amount supplied by all firms besides firm i, and the term $\frac{\partial Q_{-it}}{\partial q_{it}}$ is referred to as firm i's conjectural variation about its competitors' responses to a one unit change in firm i's output level.

Although we discuss conjectural parameters in more detail in the next section, one way to think about the conjectural variation parameter is that it indexes how far price is from marginal cost. If the firm chooses its output assuming it has no influence on market price, then it perceives that any increase in output will be met with an equal and opposite change in the aggregate output of its competitors so that market prices are unchanged. This means $\frac{\partial Q_{-it}}{\partial q_{it}} = -1$, so that θ_{it} equals zero and price equals marginal cost, which implies that the firm assumes it is unable to affect the market price through its quantity-setting actions. For static Cournot–Nash competitors, $\frac{\partial Q_{-it}}{\partial q_{it}} = 0$, which implies that θ_{it} equals firm i's quantity share of the market. For a quantity or price-setting monopoly or cartel, the firm knows that all firms will respond one-for-one with its output change from their current level of output, so that $\frac{\partial Q_{-it}}{\partial q_{it}} = \frac{Q_{-it}}{q_{it}}$, and θ_{it} equals one. This value of θ_{it} implies monopoly pricing on the part of the cartel. Although in principle conjectural variation parameters can continuously range between zero and one, it is unclear what behavioral meaning one would attach to all other values of θ_{it} in this interval besides the three values described above.

While Porter's economic model applies to individual firm decisions, he chooses not to estimate firm-level models. This decision appears to be made because estimating firm-level specifications would add significantly to his computations, particularly if he estimated conjectural variation and cost parameters for each firm. Given the state of computing power at the time he estimated his model, we doubt this would have been computationally feasible. Additionally, such an approach would require him to model new entry during the sample period.

As is common when only industry-level price and quantity data are available, Porter instead aggregates the firm-level first-order conditions to obtain an industry supply equation of the form (37). This approach, while reducing the number of estimating equations, is not without limitations. In aggregating the first-order conditions, it quickly becomes clear that one cannot estimate separate conjectural and cost parameters for each firm and time period. To reduce the dimensionality of the parameters in the industry supply function, Porter assumes that the firm-level values of θ_{it} times the associated market shares are the same (unknown) constant. This assumption has the important computational advantage of reducing the number of conjectural and cost parameters to two. Moreover, it makes it easy to calculate equilibrium prices and quantities in perfectly competitive and monopoly (collusive) markets. It should not be surprising that this simplifying assumption has disadvantages. The two main ones are that the model is now inconsistent with a Cournot market outcome and it is unclear why conjectural parameters should vary inversely with market shares.

Porter obtains his supply equation by weighting each firm's first-order condition in (40) by its quantity,

$$p_t\left(1 + \frac{\theta_t}{\epsilon}\right) = DQ_t^{\delta-1}, \tag{41}$$

where

$$D = \delta\left(\sum_{i=1}^{N} a_i^{\frac{1}{1-\delta}}\right)^{1-\delta}, \tag{42}$$

$$\theta_t = \sum_{i=1}^{N} s_{it}\theta_{it}, \tag{43}$$

and $s_{it} = q_{it}/Q_t$ is the quantity share of firm i in time t. Taking the natural log of this equation yields the aggregate supply function that Porter estimates, apart from the addition of an error term.

At this point, it is useful to summarize Porter's structural model. The main attraction of Porter's assumptions are that they result in a two-equation linear (in the parameters) system that explains equilibrium industry price and quantity data:

$$\begin{aligned} &\ln Q_t - \epsilon \ln p_t = \alpha + Z_t\gamma + \nu_t && \text{Demand Equation,} \\ &-(\delta - 1)\ln(Q_t) + \ln p_t = \lambda + \beta I_t + W_t\phi + \eta_t && \text{Supply Equation,} \end{aligned} \tag{44}$$

where $\lambda = \ln D$, $\beta = -\ln(1 + \theta/\epsilon)$, I_t is an indicator random variable which takes on the value 1 when the industry is in a cooperative regime and 0 when the industry is in a competitive regime, W_t is a set of explanatory variables that capture aggregate supply shifts due to such events as the entry of new firms, and β is an unknown parameter that measures the extent to which price and quantities sold during the collusive regime approach the joint profit-maximizing monopoly solution. For example, if $\beta = -\ln(1 + 1/\epsilon)$, the collusive regime involves joint profit maximization. Lower values of β, however, imply higher output in the collusive regime. Porter argues based on his work with Green, that the true β should be less than the joint profit-maximizing value.

5.2.3. The stochastic model

Porter completes the economic model above with two sets of stochastic assumptions. The first set is fairly standard: he assumes the errors in the demand and industry supply equations are additive, mean zero, homoscedastic normal errors. The source of these errors is left unspecified. One presumes that each error represents demand and cost factors unobservable to modern researchers, but observable to the firms at the time. Porter also assumes the demand and supply errors are independent of the right-hand side exogenous variables. By inspection of the aggregated first-order conditions for profit-maximization in Equation (41), we can see that the supply shock can be rationalized as a common multiplicative supply shock to all firms' variable cost functions. For example, if we redefine a_i in the variable cost function for firm i as $\alpha_{it} = a_i \exp(\eta_t)$, then solving the first-order conditions for each firm and solving for the aggregate supply function, would yield supply functions with the stochastic shock, η_t, given above.

The second stochastic specification Porter adds is less conventional and is motivated by an identification problem. In principle, Porter would like to use data on I_t, which indicates when the cartel was effective, to estimate β (and thereby recover the price–cost markup parameter θ). Unfortunately, he has incomplete historical information on when the cartel was effective. Although he uses some of this information to compare prices and evaluate his model *ex post*, in his main estimations he treats I_t as a random variable that is observable to the firms but not to him. Thus, in effect the error term in the supply equation becomes $\beta I_t + \eta_t$. Absent further information on I_t, it is clear that we have an identification problem – we cannot separately recover the key parameters θ and λ. This problem is akin to having two constant terms in the same regression. To see the problem, notice that the expected value of the error (assuming η_t has mean zero) is $\beta E(I_t)$. This expectation is by assumption nonzero because $E(I_t)$ is the expected value of I_t, which equals the probability that the firms are colluding. Assuming this probability does not change over the sample, which is assumed by Porter's formulation, the nonzero average error is absorbed into the supply equation's constant term, giving $\lambda + E(I_t) = \lambda + \beta\tau$, where τ equals the probability that I_t equals one. The supply disturbance becomes $\beta(I_t - \tau) + \eta_t$. As we can see from the constant term, even if we know the constant β, we cannot separately estimate λ and τ.

To gain another perspective on identification issues in Porter's model, it is useful to compare Porter's model to the linear demand and supply model (3), discussed in the previous section. Porter's demand and supply system has the form

$$
\begin{aligned}
&y_t'\Gamma + x_t'B = E_t', \qquad (45)\\
&\begin{bmatrix} \ln Q_t & \ln p_t \end{bmatrix}\begin{bmatrix} 1 & -(1-\delta) \\ -\epsilon & 1 \end{bmatrix} + \begin{bmatrix} 1 & Z_t' & W_t' \end{bmatrix}\begin{bmatrix} -\alpha & -(\lambda+\beta\tau) \\ -\gamma & 0 \\ 0 & -\phi \end{bmatrix}\\
&\quad = [\nu_t, \beta(I_t - \tau) + \eta_t].
\end{aligned}
$$

Given the parallel, we might be tempted to use the assumptions we applied there, namely that Z_t and W_t are uncorrelated with $\beta(I_t - \tau) + \eta_t$ and ν_t and that the disturbances have a constant covariance matrix. Under these assumptions, we could obtain consistent estimates of the structural parameters, Γ, B and

$$
E\big(E_t E_t'\big) = \Sigma^* = \begin{bmatrix} E(\nu_t^2) & E(\nu_t\eta_t) \\ E(\nu_t\eta_t) & E[(\beta(I_t-\tau)+\eta_t)^2] \end{bmatrix}
$$

in Equation (45) by three-stage least squares.

Notice, however, that in the above formulation, the regime-shift variable I_t only appears in the error term. This suggests that in order to distinguish between Porter's regime-switching model and the classical model, we need to rely on the distributional assumptions Porter imposes on I_t, η_t and ν_t. Absent specific distributional assumptions for I_t and η_t, we have no hope of estimating the probability of regime shifts, τ, or the magnitude of the conduct parameter during these collusive regimes, θ, which is a nonlinear function of β, from the joint distribution of price and quantity data. To identify these parameters, Porter needs to add assumptions. This should not be too surprising given

that he does not observe I_t. His strategy for achieving identification is to parameterize the distribution of the unobservable regimes. Specifically, he assumes that I_t follows an independent and identically distributed (i.i.d.) Bernoulli process and that the ν_t and η_t are i.i.d. jointly normally distributed errors. Further, Porter assumes the demand and supply errors are independent of I_t. These assumptions allow him to identify λ and β separately.

The advantage of Porter's structural framework is that we can explore how these assumptions facilitate identification and estimation. By modeling I_t as an unobservable Bernoulli, Porter has introduced a nonnormality into the distribution of the structural model's errors. To see this, notice that conditional on the regime, the second element of E_t possesses a symmetric normal distribution. Unconditionally, however, the distribution of E_t now is composed of a (centered) Bernoulli and a normal random variable. Consequently, unlike the traditional demand and supply model (3) where we could use standard instrumental variables to recover the relevant structural parameters from conditional mean functions, here we must use more information about the joint distribution of prices and quantities to estimate the model parameters. Put another way, it is the nonnormality of the reduced form errors that determines the extent to which one can identify β empirically. This then raises the delicate question: How comfortable are we with the assumption that η_t and ν_t are normally distributed? Unless there is a compelling economic reason for assuming normality, we have to regard (as Porter does) any inference about regime shifts as potentially hinging critically on this maintained assumption. Fortunately, in Porter's case he does have some regime classification data from Ulen (1978) that agrees with his model's classification of regimes.

At this point it is useful to recall our notion of structure in a simultaneous equations model. As discussed in Section 3, the most that can be identified from descriptive analysis is the conditional density of the endogenous variables, $y_t = (\ln p_t, \ln Q_t')$, given the vector of exogenous variables, $x_t = (1, W_t', Z_t')'$; that is, $f(y_t \mid x_t)$. According to Porter's theoretical model, this observed conditional density is the result of the interaction of industry demand and an industry 'supply' that switches between collusive and noncooperative regimes. However, no amount of data will allow the researcher to distinguish between this regime-switching structural model and a conventional linear simultaneous equations model with no regime switching.

To derive the likelihood function for the case of a single regime linear simultaneous equation model, consider the error vector in Equation (45). The first error is by assumption a mean-zero normal random variable and the second is the sum of a centered Bernoulli random variable, $I_t - \tau$ and a mean zero normal random variable. Applying the law of total probability formula yields the following density for E_t

$$g(E_t) = \tau \frac{1}{2\pi} |\Sigma|^{-1/2} \exp\left(-\frac{F_{1t}' \Sigma^{-1} F_{1t}}{2}\right) + (1-\tau) \frac{1}{2\pi} |\Sigma|^{-1/2} \exp\left(-\frac{F_{2t}' \Sigma^{-1} F_{2t}}{2}\right),$$

where

$$F_{1t} = \begin{bmatrix} E_{1t} \\ E_{2t} - \beta(1-\tau) \end{bmatrix} \quad \text{and} \quad F_{2t} = \begin{bmatrix} E_{1t} \\ E_{2t} + \beta\tau \end{bmatrix}.$$

Both models give rise to the same conditional density $f(y_t \mid x_t)$, but have very different economic implications. The first model implies random switches from competitive to collusive pricing regimes; the other implies a single-pricing regime but a nonnormal distribution for E_t. Consequently, any test for regime shifts must be conditional on the assumed supply and demand functions, and more importantly, the assumed distributions for I_t and η_t. Because these distributional assumptions are untestable, as this example illustrates, we believe that any test for stochastic regime shifts, should be interpreted with caution.

One might view this result as a criticism of structural modeling. To do so would miss our earlier points about the strengths of a structural model. In particular, a key strength of a structural model is that it permits other researchers to ask how the modeler's assumptions may affect results. This example also illustrates our earlier meta-theorem that: absent assumptions about the economic model generating the observed data, the researcher can only describe the properties of the joint distribution of x_t and y_t.

To understand all of the implications of this point, we re-write Porter's regime switching model as

$$y_t'\Gamma = x_t'D + I_t\Delta + U_t', \tag{46}$$

where

$$\Gamma = \begin{bmatrix} 1 & -(1-\delta) \\ -\epsilon & 1 \end{bmatrix}, \quad \Delta = [\,0 \quad \beta\,], \quad D = \begin{bmatrix} \alpha & \lambda \\ \gamma & 0 \\ 0 & \phi \end{bmatrix}, \quad U_t = \begin{bmatrix} \nu_t \\ \eta_t \end{bmatrix},$$

$$\text{and} \quad U_t \sim N(0, \Sigma). \tag{47}$$

In terms of this notation, the conditional density of y_t given x_t and I_t is:

$$h(y_t \mid I_t, x_t) = \frac{1}{2\pi}|\Sigma|^{-1/2} \times \exp\left(-\frac{(y_t'\Gamma - x_t'D - I_t\Delta)\Sigma^{-1}(y_t'\Gamma - x_t'D - I_t\Delta)'}{2}\right).$$

Using the assumption that I_t is an i.i.d. Bernoulli random variable distributed independent of U_t and x_t yields the following conditional density of y_t given x_t:

$$f(y_t \mid x_t) = \tau h(y_t \mid I_t = 1, x_t) + (1-\tau)h(y_t \mid I_t = 0, x_t).$$

As has been emphasized above and in Section 3, all that can be estimated from a statistical analysis of observations on x_t and y_t is the true joint density of $f^{\text{true}}(y_t, x_t)$, from which one can derive the conditional density of y_t given x_t. The fact that $f^{\text{true}}(y_t \mid x_t)$, the true conditional density, can be factored into the product of two conditional normal densities times the probability of the associated value of I_t is due solely to the

functional form and distributional assumptions underlying Porter's stochastic economic model.

Without imposing this economic structure on $f(y_t \mid x_t)$, the researcher would be unable to estimate underlying economic primitives such as the price elasticity of demand, the price elasticity of supply, the probability of a collusive versus a competitive regime, and the magnitude of the difference in prices between the collusive and competitive regimes. Even the best descriptive analysis would yield little useful economic information if the true data-generation process was Porter's structural model. Suppose that one had sufficient data to obtain a precise estimate of $f^{\text{true}}(y_t, x_t)$ using the techniques in Silverman (1986). From this estimate, the researcher could compute an estimate of $E(y_t \mid x_t)$ or the conditional density of y_t given x_t. However, suppose the researcher computed $\frac{\partial E(y_t|x_t)}{\partial x_{it}}$ for the ith element of x_t. If Porter's model were correct, this expectation would equal

$$\tau \frac{\partial E(y_t \mid I_t = 1, x_t)}{\partial x_{it}} + (1-\tau)\frac{\partial E(y_t \mid I_t = 0, x_t)}{\partial x_{it}},$$

so that any partial derivative of the conditional mean is an unknown weighted sum of partial derivatives of the conditional means under the competitive and collusive regimes. The researcher would therefore have a difficult time examining the validity of comparative statics predictions concerning signs of these partial derivatives under competition versus collusion, unless the sign predictions were the same under both regimes. Inferring magnitudes of the competitive or collusive comparative static effects, would be impossible without additional information.

This last observation raises an important point about the success we would have in trying to enrich the economic model of regime shifts. Imagine, as some have, that there are more than two regimes. We might attempt to model this possibility by assuming that I_t has multiple points of support. This seemingly more reasonable model imposes greater demands on the data, as now the extent to which these additional supply regimes are "identified" is determined by a more complicated nonnormal structure of the reduced form errors.

One final point about the estimation of β is that care must be exercised in drawing inferences about the presence of multiple regimes. Under the null hypothesis that there are no regime shifts, standard likelihood ratio tests are invalid. The problem that arises is that under the null of no regime shifts, τ, the probability of the collusive regime, is equal to zero and β is no longer identified. Technically this causes problems because the information matrix is singular when $\tau = 0$. It is unclear then what meaning we can attach to standard tests of the hypothesis that there are distinct regimes.

5.3. Summary

Our analysis of Porter's model leads us to conclude that demand and supply models for oligopolistic industries pose special identification and applied econometric problems. More importantly, the parameters describing competitive conjectures or the degree of

competition are not necessarily identified with commonly available data. In general, the researcher will have to have within-sample variation in demand or cost parameters, or make specific distributional assumptions and apply specific estimation techniques, to identify how competitive conduct affects industry supply behavior. As we shall see, this identification problem is all too common in industrial organization models of firm and industry behavior.

The strength of Porter's model is that it both illustrates potential identification and estimation problems posed by the standard theory and commonly available industry data. It also provides a strategy for recovering information about competitive regimes from limited information about the prevailing competitive regime. Although one could consider alternative strategies for identifying the competitive regimes, Porter compares his estimates of the probability of collusion to information from Ulen (1978) on when the cartel was actually effective. This is an example of how other evidence can be brought to bear to check whether the results of the structural model make sense. Porter finds a remarkable amount of agreement between the two measures. His model also provides an economically plausible explanation for the enormous variation in grain prices over his sample period.

6. Market power models more generally

Porter's model is an example of an IO model that uses data on market-clearing prices and outputs to draw inferences about the extent of market competition. Because these are among the most widely used empirical models in industrial organization, it is worth going beyond Porter's model to consider what other studies have done to identify market power. There are an enormous number of market power studies, many more than we can do justice to here. Bresnahan (1989) surveys the early papers in this area. Our focus is on illustrating the critical modeling issues that arise in the identification and estimation of these models.

Most empirical researchers in IO define a competitive market outcome as one where price equals the marginal cost of the highest cost unit supplied to the market. If the market price is above this marginal cost, then firms are said to exercise "market power". While some studies are content simply to estimate price–cost margins, many go further and attempt to infer what types of firm behavior ("conduct") are associated with prices that exceed marginal costs. A first observation we make below is: absent a structural model, one cannot infer the extent of competition from the joint distribution of market-clearing prices and quantities. Put another way, one needs an economic model to estimate marginal costs (and hence price–cost margins) from the joint distribution of market-clearing prices and quantities. This structural model will involve functional form assumptions and often distributional assumptions that cannot be tested independently of hypotheses about competition.

While this observation may seem obvious from our discussion of Porter's model, there are plenty of examples in the literature where researchers draw unconditional in-

ferences about the extent of competition. That is, they draw inferences about price–cost margins without acknowledging that their inferences may depend critically on their economic and functional form assumptions.

A second observation below is: while one can estimate price–cost margins using a structural model, it is problematic to link these margins to more than a few specific models of firm behavior. In particular, many studies estimate a continuous-valued parameter that they claim represents firm "conjectures" about how competitors will react in equilibrium. Currently there is no satisfactory economic interpretation of this parameter as a measure of firm behavior – save for firms in perfectly competitive, monopoly, Cournot–Nash and a few other special markets. We therefore see little or no value to drawing economic inferences about firm conduct from conjectural variation parameter estimates.

In what follows we discuss these two observations in more detail. We first discuss how the literature identifies and interprets market power within the confines of static, homogenous goods models where firms choose quantities. We then discuss at a broader level what market power models can tell us in differentiated product markets.

6.1. Estimating price–cost margins

Since the late 1970s, many papers in IO have used firm and industry price and quantity data to describe competition in homogeneous product markets. The typical paper begins, as Porter did, by specifying a demand function and writing down the first-order condition:

$$P + \theta_i q_i \frac{\partial P}{\partial Q} = \mathrm{MC}_i(q_i). \tag{48}$$

The goal of these papers is to estimate the 'conduct' parameter θ_i. Most authors assert that this parameter measures firm "conjectures" about competitor behavior. As such, it would seem to be a structural parameter that comes from an economic theory. Is this the case?

Isolating θ_i in Equation (48), and letting α_i denote firm i's output share and ϵ the elasticity of demand, we obtain

$$\theta_i = \frac{P - \mathrm{MC}_i(q_i)}{-q_i \frac{\partial P}{\partial Q}} = \frac{P - \mathrm{MC}_i(q_i)}{P} \frac{1}{\alpha_i \epsilon}. \tag{49}$$

From this equation, we see that θ_i provides essentially the same *descriptive* information as Lerner's (1934) index. That is, it provides an idea of how far a firm's price is from its marginal cost. To the extent that price is above marginal cost (i.e., the Lerner index is positive), IO economists claim that the firm has 'market power'.

Equation (49) is useful because it identifies two critical structural quantities that a researcher must have to estimate θ_i. These are the price elasticity of demand and marginal cost. Following Porter, a researcher could in principle separately estimate the price elasticity of demand from price and quantity data. In developing such an estimate, the

researcher would of course have to worry that the demand function's form may critically impact the estimated elasticity. The marginal cost term in Equation (49) poses a more difficult estimation problem. Equation (49) tells us that with just price and quantity data, we cannot separate the estimation of marginal cost from the estimation of θ_i. Even if we have observations on total or even variable cost associated with this level of output, we are unable to separate them without making specific functional form assumptions for demand and marginal cost. Put another way, the identification of θ_i hinges on how we choose to estimate marginal cost and the aggregate demand curve. Changing the marginal cost and demand specification will change our estimate of θ_i. Unless one knows the functional form of demand and costs, it is impossible to determine the value of θ_i.

Despite the many untestable functional form assumptions necessary to infer marginal costs from price and quantity data, many studies go further and use Equation (48) to estimate θ_i and interpret it as a measure of firm behavior. To understand where this behavioral interpretation comes from, we return to the economic rationale underlying Equation (48). In Equation (48), θ_i is a placeholder for the derivative:

$$\theta_i = \frac{dQ}{dq_i}. \tag{50}$$

According to this definition, θ_i is not a statement about how far prices are from marginal costs, but rather a "variational" concept associated with firm behavior. Specifically, Equation (48) sometimes is interpreted as saying: the firm "conjectures" industry output will increase by θ_i should it increase its output by one unit. The problem with this interpretation is that there are only a few values of θ_i where economists have a good explanation for how firms arrived at such a conjecture. This leads to our second observation above. We know of no satisfactory static model that allows for arbitrary values of θ_i. Empirical models that treat θ_i as a continuous value to be estimated thus are on shaky economic ground, particularly because these estimates of θ_i are predicated on a specific functional form for marginal costs and demand.

To emphasize the danger inherent in associating residually determined θ_i with behavior, imagine observing two firms producing different quantities who otherwise appear identical. The conjectural variation approach would explain the difference by saying firms simply "expect" or "conjecture" that their competitors will react differently to a change in output. Yet there is no supporting story for how otherwise firms arrived at these different conjectures. On the other hand, even though the firms appear identical, one might wonder whether their marginal costs are identical. It seems plausible to us that unobservable differences in marginal costs, rather than behavior, could explain the difference in output. Absent a richer model of behavior that explains where conjectures come from, it is anyone's guess.

To summarize our discussion so far, we have provided two possible interpretations of θ_i. There are, however, a few instances in which θ_i sensibly corresponds to a specific market equilibrium. A leading case is price-taking competition, where $\theta_i = 0$ and price equals marginal cost. Cournot ($\theta_i = 1$), Stackleberg and monopoly are three other

well-known cases. While there has been some debate in the theoretical literature about whether these models are internally "consistent" static behavioral models [e.g., Lindh (1992)], each of these models lends itself to a natural interpretation of what θ_i means as a conjecture about competitor behavior. Thus, it seems to us sensible to imagine imposing these conjectures in the first-order condition (48) and using them to estimate the parameters of demand and cost functions. One can use nonnested tests, as in Bresnahan (1987), to determine which of these different models of behavior are most consistent with the joint density of the data.

Having said this, we realize that some might argue that one loses little by treating θ_i as a continuous parameter to be estimated. After estimating it, the argument goes, one can still compare it to the benchmark values. For example, suppose one precisely estimated $\theta_i = 1.7$, and could reject perfect competition and Cournot. One might think it reasonable to conclude the market is "less competitive than Cournot". But does this make much sense? According to the conjectural variations story, and Equation (48), an estimate of 1.7 implies that firm i believes that if it increases output by one unit, industry output will increase by 1.7 units. What type of behavior or expectations leads to firm i maximizing its profits by maintaining $\theta_i = 1.7$? Why does this value not simply reflect the extent of misspecification of the demand and cost functions in a Cournot model? The problem here is that the theory underlying firm i's behavior (and those of its competitors' behavior) is static. There is no obvious explanation for why firm i has this behavior. Moreover, as we show in the next section, in order to identify an estimate of θ_i, a researcher must select a parametric aggregate demand curve and rule out several types of functional forms for aggregate demand. Otherwise it is impossible to identify θ_i from market-clearing price and quantity data.

If there is an answer to the question of where a firm's conjectures comes from, it must come from a dynamic model of "conjectures" formation. Riordan (1985) provides one such model. Given the subtleties involved with reasoning through how today's competitive interactions might affect future beliefs, it seems unlikely dynamic models will produce simple parameterizations of conjectures or easily estimated first-order conditions. Moreover, the literature on repeated games has shown that when modeling current behavior, one has to recognize that threats or promises about future behavior can influence current behavior. This observation points to a distinction between what firms do in equilibrium (how they appear to "behave") and what they conjecture their competitors' would do in response to a change in each firm's output.[4] This also is a distinction that Stigler (1964) used to criticize static conjectural variation models.

To understand how this distinction affects empirical modelers, consider a cartel composed of N symmetric firms, each of whom charges the monopoly price. In this case, one would estimate θ_i equal to the number of firms. If we gave this estimate a behavioral interpretation, we would report that in this industry, firms conjecture or expect other firms to change their outputs one-for-one. Yet this may not be the case at all, as

[4] Corts (1999) makes a similar argument.

some recent theories have emphasized. The firms may be charging the monopoly price because they expect that if they defect from the monopoly price by producing a little more, each of their competitors may punish them by producing much more.

This distinction between the "beliefs" that economic agents hold and what they ultimately may do in equilibrium is critical for exactly the reasons we outlined in our introductory framework. If one wants to describe where price is in relation to a firm's marginal cost, then θ_i provides a descriptive measure of that, but not a statement about behavior. If, however, one wants to use the estimated parameters to predict what would happen if the firms' economic environment changes, then one either must have a theory in which beliefs and equilibrium behavior coincide, or one must ask which of a small set of values of θ_i, corresponding to perfect competition, monopoly, Cournot and the like, best explains the data.

6.2. *Identifying and interpreting price–cost margins*

In the previous subsection we emphasized that while one could relate θ to price–cost margins, one could not separately estimate θ and marginal costs from price and quantity data alone. Despite occasional claims to the contrary, assumptions about the functional form of marginal costs are likely to affect estimates of θ and vice versa. This section illustrates how assumptions about the structure of demand and marginal costs impact the estimation of the descriptive parameter θ. (Throughout this subsection, we think of θ as providing descriptive information about price–cost margins.)

The IO literature has adopted different approaches to estimating price–cost margins depending upon whether or not they have individual firm or industry price and quantity data. When only industry-level data are available, researchers typically use the equation

$$P + \theta Q \frac{\partial P}{\partial Q} = \mathrm{MC}(Q) \tag{51}$$

to estimate a single industry θ. James Rosse's (1970) paper is the first to estimate the degree of market power (the price–cost markup), or equivalently a firm's marginal cost curve. He used observations on market-clearing prices and quantities from a cross-section of US monopoly newspaper markets. Rosse's procedure uses this first-order condition with θ set equal to 1, along with an assumed parametric aggregate demand curve to estimate the marginal cost curve. This procedure works for the following reason. Once a parametric functional form for demand is selected, this can be used to compute $\frac{\partial P}{\partial Q}$ for each observation in the sample. Setting the value of θ for each observation to 1 guarantees that we have the information necessary to compute the left-hand side of Equation (51) for each observation. This provides an implied value of marginal cost for every output level in the sample. Combining this data with a parametric specification for the firm's marginal cost function, we can estimate marginal cost parameters.

To extend Equation (51) to an oligopoly market requires further assumptions. This equation would appear to mimic a single firm's first-order condition, and thus we might

think of it as linked to the price–cost margins of a "representative" firm. But this is not generally true. Starting as Porter did from the individual firm profit maximization conditions, we can sum Equation (48) across firms to obtain the relation

$$P + \frac{\partial P}{\partial Q} \sum_{i=1}^{N} \frac{\theta_i q_i}{N} = \sum_{i=1}^{N} \frac{\mathrm{MC}(q_i)}{N}, \tag{52}$$

which we can rewrite as

$$P + \theta \frac{\partial P}{\partial Q} Q = \overline{\mathrm{MC}(q_i)}. \tag{53}$$

Here, $\theta = \frac{1}{N} \sum_{i=1}^{N} \frac{\theta_i q_i}{Q}$ is an average of firm market shares times the individual firm θ_i parameters, and $\overline{\mathrm{MC}(q_i)}$ is the average of the N firms' marginal costs. While this equation "looks" like the industry aggregate equation (51) used in many studies, it is not the same without further assumptions. Note, for example, that if θ_i varies across firms, then changes in firms' market shares will generally change θ. Thus, if one is analyzing time series data on prices and output, it may make little sense to treat θ in Equation (51) as a constant. An exception is when one assumes all firms have the same θ_i. But in this case, one must have the same number of firms in the industry for θ to remain constant through time.

The assumption that all firms have the same θ_i amounts to assuming that at the same production level, all firms in the industry would have similarly sloped firm-level demand curves and the same marginal revenues. This is a nontrivial restriction which would require justification on a case-by-case basis. A number of studies, beginning with Gollop and Roberts (1979), Applebaum (1982) and Spiller and Favaro (1984), have argued that one should relax this restriction by making θ a function of different variables, including output. To date, however, there is very little economic theory to guide structural models of how θ_i varies across firms. The most widely adopted specifications are ad hoc, with θ depending on firm output, market share or a firm's size rank.

Another consequence of assuming all firms have the same θ is that differences in firms' outputs now are a function solely of differences in marginal costs. In some instances, this leads to a monotonic relationship between the efficiency of a firm and its observed production. For example, if we assume marginal costs are increasing in output, then there is an inverse relationship between output and marginal costs. Thus, the firm with the largest output has the lowest marginal cost, the firm with the second largest output the second lowest marginal cost, and so on. While this relationship may be entirely reasonable for many industries, it may not be for all.

Turning now to the right-hand side of Equation (51), we see that the notation $\mathrm{MC}(Q)$ gives the impression that only industry output enters the industry supply relation. Put another way, a reallocation of output from one firm in the industry to another will not change the right-hand side of the industry supply relation (51). This obviously cannot generally be true. Equation (53) shows why this is so. To explore this point further, it is

useful to assume that firms have linear marginal costs of the form

$$\mathrm{MC}(q_i) = c_{0i} + c_{1i} q_i. \tag{54}$$

In this case, we can rewrite Equation (53) as

$$P + \tilde{\theta} Q \frac{\partial P}{\partial Q} = \bar{c_o} + \tilde{c_1} Q + \psi, \tag{55}$$

where

$$\tilde{\theta} = \frac{\sum_{i=1}^{N} \frac{\theta_i}{N}}{N}, \tag{56}$$

$$\bar{c_0} = \frac{1}{N} \sum_{i=1}^{N} c_{0i}, \qquad \tilde{c_1} = \frac{1}{N^2} \sum_{i=1}^{N} c_{1i}, \tag{57}$$

$$\psi = \mathrm{Cov}(c_{1i}, q_i) - \mathrm{Cov}(\theta_i, q_i) \frac{\partial P}{\partial Q} \tag{58}$$

and $\mathrm{Cov}(x, y)$ equals the covariance (calculated over firms in the industry) between x and y. If the ψ term is zero, then Equations (53) and (51) are indistinguishable. This happens for example when firms have similarly sloped marginal cost functions and the same θ. In general, however, we can think of Equation (51) as having an error term that includes ψ. To the extent that ψ is nonzero and varies systematically in the researcher's sample, the researcher will obtain biased estimates of the demand, cost and θ parameters by ignoring ψ.

We now turn to considering whether and how functional form assumptions might affect inferences about θ from industry price and quantity data. Both Bresnahan (1982) and Lau (1982) consider the issue of identification in detail using the aggregate equation (51). Because their results apply to a special aggregation of individual firm first-order conditions, it is useful to revisit their discussion in the context of the individual firm marginal revenue equal to marginal cost conditions. To facilitate this discussion, let each firm face the demand function $Q = D(P, Y, \alpha)$, where α is a vector of demand parameters and Y is a set of exogenous variables that shift demand but not cost. Suppose also that each firm has the marginal cost function $\mathrm{MC}_i = c_0 + c_1 q_i + c_2 w_i$, where w_i is an exogenous cost shifter. If a researcher had time series data on market prices, firm i's output, Y and w_i over time, the researcher could estimate firm i's market power parameter θ_i using the two equation system

$$\begin{aligned} Q &= D(P, Y, \alpha), \\ P &= c_0 + \left(c_1 + \frac{\partial D^{-1}}{\partial Q} \theta_i \right) q_i + c_2 w_i \end{aligned} \tag{59}$$

once some assumption had been made about unobservables. The second equation shows that by assuming marginal costs are linear in output, we have potentially destroyed the identification of θ_i. Consider, for example, what happens when demand has the form

$Q = \alpha_0 + \alpha_1 P + \alpha_2 Y$. In this case, firm i's supply relation is

$$P = c_0 + \left(c_1 + \frac{\theta_i}{\alpha_1}\right) q_i + c_2 w_i. \tag{60}$$

Hence, even though we can obtain a consistent estimate of the demand parameter α_1 from the demand equation, we cannot separate c_1 from a constant θ_i. Of course, if we are willing to restrict θ, we can identify the marginal cost parameters and price–cost margins.

It is tempting to identify θ_i in this case by assuming that marginal costs are constant, i.e., $c_1 = 0$. Unfortunately, researchers rarely have independent information that would support this assumption. Alternatively, following Bresnahan (1982), one could identify θ_i by allowing the slope of market demand to vary over time in an observable way. For instance, one might interact price with income (Y) in the demand equation

$$Q = \alpha_0 + \alpha_1 P + \alpha_2 Y P$$

to obtain the supply equation

$$P = c_0 + \left(c_1 + \frac{\theta_i}{\alpha_1 + \alpha_2 Y}\right) q_i + c_2 w_i. \tag{61}$$

Although θ_i is formally identified in this specification, its identification in practice depends heavily on having variables, such as income, that interact or otherwise cannot be separated from price [e.g., Lau (1982)]. In other words, the value of θ is identified off of a functional form assumption for aggregate demand.

Yet another approach to identifying θ_i that has not been fully explored is to add information from other firms' supply relations. In the language of econometrics, it may be possible to obtain identification by imposing cross-equation restrictions between the pricing equations. Returning to the specification in Equation (53), if we added a supply curve for a second firm j, we still would not be able to identify θ_i or θ_j. We would, however, be able to identify the difference if we assumed that both firms' marginal cost functions had the same slope. Alternatively, we could identify the difference in the slopes of the firms' marginal cost functions if in a panel data setting (where T goes to infinity) we assume that all firms have the same constant θ.

Our discussion so far has suggested that θ is identified by the functional form assumptions one makes about market demand and firms' costs. This dependence seems to not always be appreciated in the literature, where cost and demand functions are sometimes written down without much discussion of how their structure might affect estimates of θ. A useful example of how the functional form of demand affects the identification of θ is provided by the inverse demand function:

$$P = \alpha - \beta Q^{1/\gamma}. \tag{62}$$

This inverse demand function leads to the direct estimator (by applying Equation (49) above)

$$\theta_1 = -\gamma \frac{P - c}{\alpha - P}, \tag{63}$$

which illustrates how the demand parameters affect the direct estimate. This inverse demand function also yields a transformed Equation (51)

$$P_t = \frac{\gamma c_t}{\gamma + \theta} + \frac{\alpha\theta}{\gamma + \theta}, \tag{64}$$

where the subscript t denotes variables that are naturally thought of as time varying. Critical to most applications is what one assumes about marginal costs. In the simplest case, one can think of firms as having constant, but time-varying marginal costs c_t which depend linearly on some time-varying exogenous covariates, i.e.,

$$c_t = c_0 + W_t\omega,$$

where ω is a vector of parameters. Substitution of this relationship into (64) gives the equation

$$P_t = \frac{\alpha\theta}{\gamma + \theta} + \frac{\gamma c_0}{\gamma + \theta} + \frac{\gamma}{\gamma + \theta} W_t\omega = \beta_0 + \beta_1 W_t.$$

This equation makes it clear that absent further assumptions, we cannot identify θ from estimates of β_0 and β_1 alone. One way around this problem is to recognize from Equation (53) that θ depends on market shares and the number of firms, both of which are potentially time varying. This, however, is not the usual approach. Instead, most studies follow the advice of Bresnahan and Lau and identify θ by assuming that the demand parameters α and/or γ contain a demand covariate. For example, if we assume that the inverse demand intercept equals

$$\alpha_t = \alpha_0 + D_t\alpha_1,$$

then Equation (64) becomes

$$P_t = \frac{\alpha_0\theta}{\gamma + \theta} + \frac{\gamma c_0}{\gamma + \theta} + \frac{\alpha_1\theta}{\gamma + \theta} D_t + \frac{\gamma}{\gamma + \theta} W_t\omega.$$

This equation and the demand equation now exactly identify θ. But note that the estimate of θ depends critically on the effect of D on demand and on the curvature of demand. If we had started out, as many studies do, by assuming linear demand then we could not estimate θ. This is yet another example of how economic structure is solely identified by a functional form or distributional assumption.

6.3. Summary

In this section we have discussed how IO researchers use price and quantity data to estimate price–cost margins. We also have questioned the value of static conjectural variation parameters. Apart from these observations, we have tried to underscore one of the key observations of our framework, which is that functional form assumptions play a critical role in inferences about marginal economic "effects" and the appropriate model of competition.

7. Models of differentiated product competition

The previous two sections discussed how IO economists have used price and quantity data to draw inferences about the behavior of oligopolists selling homogeneous products. These empirical models parallel textbook demand and supply models. The chief difference is in an oligopoly model, the supply equation is replaced by a price equation derived from first-order conditions that describe how oligopolists maximize profits. Because IO economists do not observe the marginal costs that enter these first-order conditions, they are forced to estimate them along with other structural parameters. It should not be too surprising that a researcher's stochastic and functional form assumptions have a critical impact on the resulting estimates, as the researcher is simultaneously trying to draw inferences about the nature of demand, costs and competition from just data on prices and quantities.

This section examines how IO economists have used price and quantity information on differentiated products to draw inferences about demand, costs and competition. We first discuss complexities that arise in neoclassical extensions of homogeneous product models. We then turn to more recent differentiated product discrete choice models.

7.1. Neoclassical demand models

In the late 1980s and 1990s, empirical IO economists began to focus on modeling competition in differentiated product markets such as cars, computers and breakfast cereals. Bresnahan (1981, 1987) are two early examples of this work. These models also use price and quantity data to draw inferences about oligopolists' strategic interactions and price–cost markups. The main difference between these models and homogeneous product models is that the researcher specifies separate "demand" and "supply" equations for each product. Thus, instead of working with two-equation, market-level systems such as (36) and (37), the researcher specifies a J-product demand system:

$$\begin{aligned} Q_1^d &= h_1(P_1, P_2, \ldots, P_J, Z_1, \beta_1, \nu_1), \\ &\vdots \\ Q_J^d &= h_J(P_1, P_2, \ldots, P_J, Z_J, \beta_J, \nu_J) \end{aligned} \tag{65}$$

and a J-equation system of first-order profit maximization conditions:

$$\begin{aligned} P_1 &= g_1\big(Q_1^s, Q_2^s, \ldots, Q_J^s, W_1; \theta_1, \eta_1\big), \\ &\vdots \\ P_J &= g_J\big(Q_1^s, Q_2^s, \ldots, Q_J^s, W_J; \theta_J, \eta_J\big). \end{aligned} \tag{66}$$

Although these systems look much more complicated than the simultaneous equations in the homogenous product case, they pose the same basic modeling issue: unless the researcher is willing to make specific functional form assumptions for firms' demands

and costs, the researcher will be unable to draw inferences about equilibrium firm-level markups. This issue arises again because, absent economic assumptions about the structure of demand and costs, the most the researcher can do is use flexible data-smoothing techniques to recover the conditional joint density of the J prices and J quantities given the demand and cost variables W and Z. Only by making functional form assumptions for demand and costs, and assumptions about the forms of strategic interactions, can the researcher recover information about demand and cost primitives. This means that we still need specific functional form assumptions to use price and quantity data to draw inferences about equilibrium markups.

The main new issue posed by differentiated products is one of scale. Now, the researcher has to specify a set of demand functions – potentially involving dozens or hundreds of products. Left unrestricted, the number of parameters in these demand systems can easily exceed the number of observations in conventional market-level price and quantity datasets. This problem has led IO researchers to focus on how best to formulate parsimonious, yet flexible, demand systems.

To appreciate the practical issues involved, consider the challenge IO economists or antitrust authorities face in trying to assess the competitiveness of the US ready-to-eat breakfast cereal industry. Absent cost data, inferences about manufacturer and retailer price–cost margins have to be drawn from retail prices and sales. As there are over 50 major brands of cereals, a simple model would have at least 100 equations – 50 demand and 50 "supply" equations. Each equation conceivably could contain dozens of parameters. For instance, paralleling Porter's homogeneous product specification, we could assume a log-linear demand system:

$$\begin{aligned}
\ln Q_1 &= \beta_{10} + \beta_{11}\ln y + \beta_{12}\ln P_1 + \beta_{13}\ln P_2 + \cdots + \beta_{1,51}\ln P_{50} + Z_1\gamma_1 + \nu_1,\\
\ln Q_2 &= \beta_{20} + \beta_{21}\ln y + \beta_{22}\ln P_1 + \beta_{23}\ln P_2 + \cdots + \beta_{2,51}\ln P_{50} + Z_2\gamma_2 + \nu_2,\\
&\vdots\\
\ln Q_{50} &= \beta_{50,0} + \beta_{50,1}\ln y + \beta_{50,2}\ln P_1 + \beta_{50,3}\ln P_2 + \cdots\\
&\quad + \beta_{50,50}\ln P_{50} + Z_{50}\gamma_{50} + \nu_{50}.
\end{aligned} \tag{67}$$

This system has over 2600 parameters![5] Such unrestricted parameterizations easily exceed the number of observations obtainable from public sources.

Even when the researcher has large amounts of data, Equations (65) and (66) pose significant computational challenges. For instance, to use maximum likelihood, the researcher would have to work with the Jacobian of 100 demand and markup equations. Nonlinearities in the system also raise the concern that the system may not have a unique solution or a real-valued solution for all error and parameter values. Although these complications can sometimes be dealt with in estimation, they may still reappear when the researcher performs counterfactual calculations. For instance, there may be

[5] Recall that aggregate demand need not be symmetric.

no nonnegative prices or single set of prices that solve (65) and (66) for a particular counterfactual.

These econometric issues have prompted IO researchers to look for ways to simplify traditional neoclassical demand models. Many early simplifications relied on ad hoc parameter restrictions or the aggregation of products.[6] For example, to estimate (67) a researcher might constrain a product's cross-price elasticities to all be the same for all products.[7] Simplifications such as this, while computationally convenient, can unduly constrain estimates of price–cost markups.

Multi-level demand specifications provide a somewhat more flexible demand function parameterization.[8] In a multi-level demand specification, the researcher separates the demand estimation problem into several stages or levels. At the highest level, consumers are viewed as choosing how much of their budget they wish to allocate to a type of product (e.g., cereal). At the next stage, the consumer decides how much of their budget they will divide among different categories of the product (e.g., categories of cereal such as kids', adult and natural cereals). At the final stage, the consumer allocates the budget for a category among the products in that category (e.g., within kids' cereals, spending on Trix, Count Chocula, etc.).

Although multi-stage models also restrict some cross-price elasticities, they permit flexible cross-price elasticities for products within a particular product category. For example, a researcher can estimate a flexible neoclassical demand system describing the demands for kids' cereal products. Changes in the prices of products in other categories (e.g., adult cereals) will still affect the demands for kids' cereals, but only indirectly through their effect on overall kids' cereal spending. Exactly how these restrictions affect estimates of markups is as yet unclear.[9]

Other recent work in the neoclassical demand system tradition has explored reducing the number of demand parameters by constraining cross-price effects or making them depend on estimable functions of covariates.[10] Pinkse, Slade and Brett (2002), for example, constrain the coefficients entering firms' price elasticities to be functions of a small set of product attributes. While this strategy facilitates estimation and allows flexibility in own and cross-price effects, it has the disadvantage of being ad hoc. For instance, it is not clear where the list of attributes comes from or how the functional

[6] Bresnahan's (1989) Section 4 reviews early efforts. Deaton and Muellbauer (1980) provide a survey of neoclassical demand models.

[7] One utility-theoretic framework that produces this restriction is to assume that there is a representative agent with the constant elasticity of substitution utility function used in Dixit and Stiglitz (1977).

[8] See, for example, Hausman, Leonard and Zona (1994).

[9] Theoretical work, beginning with Gorman (1959), has explored the restrictions that multi-stage budgeting models place on consumer preferences, and how these restrictions affect compensated and uncompensated price effects. See for example Gorman (1970), Blackorby, Primont and Russell (1978) and Hausman, Leonard and Zona (1994). Nevo (2000) evaluates empirically the flexibility of a multi-stage model.

[10] An early example is Baker and Bresnahan (1988). They propose a "residual" demand approach which forsakes identification of the original structural parameters in favor of amalgams of structural parameters.

form of demand reflects the way consumers evaluate product attributes. [Davis (2000) discusses these and other tradeoffs.]

Besides having to grapple with how best to restrict parameters, each of the above approaches also has to address the joint determination of prices and quantities. As in homogeneous product models, the presence of right-hand side endogenous variables raises delicate identification and estimation issues. Applied researchers can most easily address identification and estimation issues in demand and mark-up systems that are linear in the parameters. In nonlinear systems, identification and estimation questions become much more complicated. For example, the implicit "reduced form" for the nonlinear (65) and (66) system:

$$\begin{aligned} Q_1 &= k_1(Z, W, \beta; \theta, \nu, \eta), \\ &\vdots \\ Q_J &= k_J(Z, W, \beta; \theta, \nu, \eta), \\ P_1 &= l_1(Z, W, \beta; \theta, \nu, \eta), \\ &\vdots \\ P_J &= l_J(Z, W, \beta; \theta, \nu, \eta) \end{aligned} \tag{68}$$

may not be available in closed form. As argued earlier, these equations also need not have a solution or a unique solution for all values of the right-hand side variables and errors.

The value of the reduced forms in (68) is that they suggest that there are many potential instruments for prices and quantities. For example, they suggest that one may be able to use other products' attributes and cost variables as instruments. Unfortunately, most IO data sets do not have product-specific or firm-specific cost information. Even when researchers do have cost information, this information is likely to be extremely highly correlated across products. The lack of good cost covariates has forced researchers to use the attributes of other products as instruments. These studies have used both the prices of other products as instruments and the nonprice attributes of other products as instruments.

Hausman (1997) is a good example of a study that uses the prices of other products as instruments. Hausman develops and estimates a multi-stage budgeting model for varieties of breakfast cereals. Because he does not have cost data for the different cereal products, and he lacks time-varying attribute data, he resorts to using cereal prices in other markets as instruments. He justifies using these prices as follows. He first supposes that the price for brand j in market m and time period t has the form

$$\ln p_{jmt} = \delta_j \ln c_{jt} + \alpha_{jm} + \nu_{jmt}, \tag{69}$$

where c_{jt} are product-specific costs that do not vary across geographic areas, the α_{jm} are time-invariant, product-city (m) specific markups, and ν_{jmt} are idiosyncratic unobserved markups. He also assumes that the ν_{jmt} are independent across markets. This

latter assumption allows him to assert that prices in other cities are correlated with a specific market's prices and uncorrelated with the unobservable markup or cost component in prices.

From the perspective of our framework, the essential questions are: What economic assumptions motivate the pricing equation (69)? Following our homogeneous-product discussion, the pricing equation (69) could represent either a markup relation obtained from a first-order profit-maximization condition or a reduced form equation arising from the solution of a model along the lines of (68). To see the problem with the former interpretation, imagine that each manufacturer j maximizes profits of one product in each market. Suppose the firm also has constant marginal costs. If it maximizes profits by choosing quantity, then the markup equations will have the additive form

$$P_j = c_j + \tau(Q_1, \ldots, Q_J),$$

where as in (48) the $\tau(\cdot)$ function contains an own-demand derivative. We can re-express this equation as (69)

$$\ln P_j = \ln c_j - \ln\bigl(1 - \tau(Q_1, \ldots, Q_J)/P_j\bigr).$$

The question then is whether the multi-level demand function specification Hausman uses leads to the second term above having the form $\alpha_j + \nu_{jt}$, where the ν's are independent across markets. In general, his flexible demand system would not appear to lead to such a specification.

One could imagine alternatively that (69) is the reduced form obtained by simultaneously solving the first-order conditions. The problem with this view is that Hausman's flexible demand specification implies the costs of all other products should enter the reduced form. This would mean that either α_{jm} would have to be time-varying to account for the time variation in other product's costs or that c_{jt} would have to be market varying.

In principle, one might imagine adjusting the multi-level demand system or the pricing equation (69) to justify using variables from other markets as instruments. Such an exercise will require additional economic and statistical assumptions. Consider, for example, the α_{jm}'s in Equation (69). These terms appear to represent unobserved product and market-specific factors that affect markups. They might, for example, capture San Franciscans' unobserved health-conscious attitudes. These attitudes might lead San Franciscans' to have a higher demand and greater willingness to pay for organic cereals. If natural cereal makers are aware of this propensity, they might advertise more in the San Francisco market. If this advertising is not captured in the demand specification, then the demand error will be correlated with the α. Hausman recognizes this possibility by removing the brand-market α's using product-market fixed effects. Letting $\widetilde{\ }$ denote the residual prices from these regressions, his results rely on the adjusted prices:

$$\widetilde{\ln p_{jnt}} = \delta_j \widetilde{\ln c_{jt}} + \widetilde{\nu_{jnt}} \tag{70}$$

as instruments. According to Equation (69), these adjusted prices would only contain adjusted national marginal costs, and residual cost and demand factors affecting

markups. At this point, Hausman still must assume that: (1) the adjusted time-varying national marginal costs $\ln c_{jt}$ are uncorrelated with the demand and cost errors in other cities; and (2) the residual demand and cost factors affecting markups are independent of the errors in other cities.

These two assumptions have been vigorously debated by Hausman (1997) and Bresnahan (1997). Bresnahan (1997) argued that there might be common unobserved seasonal factors that affect both demand and marginal costs. To illustrate this point, Bresnahan provides an example in which a periodic national advertising campaigns translate into increased demands and markups in all markets. This results in correlation between the idiosyncratic markup terms in other markets and demand errors.[11] Whether these advertising campaigns are of great consequence for demand and price–cost estimates in a particular application is not something that can be decided in the abstract. Rather it will depend on the marketing setting and the economic behavior of the firms under study.

Our discussion so far has emphasized the strong assumptions needed to use prices in other markets as instruments. Do the same arguments apply to nonprice attributes? At first, it might seem that they might not. Similar concerns, however, can be raised about nonprice instruments. Consider, for example, the problem of trying to model airline travel demand along specific city-pairs. In such a model, the researcher might use a flight's departure time as a nonprice attribute that explains demand. The reduced form expressions in (68) suggest that besides the carrier's own departure time, measures of competing carriers' departure times could serve as instruments. But what makes' the characteristics of carriers' schedules' valid instruments? They may well not be if the carriers strategically choose departure times. For example, carriers may space their departure times to soften competition and raise fares.

If firms set nonprice attributes using information unavailable to the researcher, then we can no longer be certain that product attributes are valid instruments. In some applications, researchers have defended the use of nonprice attributes with the argument that they are "predetermined". Implicit in this defense is the claim that firms find it prohibitively expensive to change nonprice attributes in the short run during which prices are set. As a result, nonprice product characteristics can reasonably be thought of as uncorrelated with short-run unobserved demand variables. For example, a researcher modeling the annual demand for new cars might argue that the size of a car is unlikely correlated with short-run changes in demand that would affect new car prices. While this logic has some appeal, it relies on the assumption that unobserved factors affecting manufacturers' initial choices of characteristics do not persist through time. This is a question that is not easily resolved in the abstract. For this reason, models that endogenize both price and nonprice attributes continue to be an active area of research in IO.

[11] The criticism that advertising influences demand amounts to an attack on demand specifications that ignore advertising. As Hausman's empirical model does include a variable measuring whether the product is on display, the question then becomes whether the display variable captures all common promotional activity.

7.2. Micro-data models

Our discussion of the product-level demand specifications in (65) has said little about what it is that leads firms to differentiate products. One ready explanation is that firms differentiate their products in order to take advantage of heterogeneities in consumer tastes. For example, car makers regularly alter a car's styling, size, drive trains and standard features to attract particular groups of buyers. If IO economists are to understand how these models are priced and compete, it seems imperative that their demand systems explicitly recognize how consumer tastes for product attributes will affect demand at the firm level. In the language of Section 4, it seems critical that *firm-level demand* models recognize both observable and unobservable heterogeneities in *individual-level tastes*. Most neoclassical demand models, however, are ill-suited to modeling consumer heterogeneities. This is because it is unwieldy to aggregate most individual-level neoclassical demand models across consumers to obtain market or firm-level demands.

In the differentiated product literature, researchers have adopted two approaches to demand aggregation and estimation. One is to estimate individual-level demand functions for a representative sample of consumers. These demand functions are then explicitly aggregated across the representative sample to obtain market or firm demand. The second is to instead assume that consumer tastes have a particular distribution in the population. This distribution, along with individual demands, are estimated together to obtain estimates of market and firm demand.

In what follows, we explore some of the advantages and disadvantages of these two approaches. To focus our discussion, we follow recent work in IO that relies on discrete choice demand specifications. These models presume that consumers buy at most one unit of one product from among J products offered.[12] While these unitary demand models are literally applicable to only a few products, such as new car purchases, they have been used by IO economists to estimate consumer demands for a range of products.

A key distinguishing feature of these discrete choice demand models is that firms are uncertain about consumers' preferences. Firms therefore set prices on the basis of expected demand. So far, firm expectations have not figured prominently in our discussion of oligopoly models. Thus, we shall begin by showing how this type of uncertainty enters a structural oligopoly model.

Imagine there are a maximum of M_t customers at time t who might buy a car. Suppose customer i has the conditional indirect utility function for car model j of

$$U_{ijt} = U(x_{jt}, p_{jt}, \omega_{ijt}),$$

where x_{jt} is a $K \times 1$ vector of nonprice attributes of car j (such as size and horsepower), p_{jt} is the car's price, and ω_{ijt} represents consumer-level variables. Consumer i will

[12] There are continuous choice multi-product demand models. These models are better termed mixed discrete continuous models because they have to recognize that consumers rarely purchase more than a few of the many products offered. See, for example, Hanemann (1984).

buy new car j provided $U(x_{jt}, p_{jt}, \omega_{ijt}) \geqslant \max_{k \neq j} U(x_{kt}, p_{kt}, \omega_{ijt}; \theta)$. If firms knew everything about consumers' tastes, they would calculate product demand as

$$\text{Demand for product } j = \sum_{i=1}^{M_t} I(i \text{ buys new car } j), \tag{71}$$

where M_t is the number of potential new car buyers at time t and $I(Arg)$ is a zero-one indicator function that is one when Arg is true. Firms would use this demand function when it came time to set prices, and the IO researcher therefore would have to do their best at approximating this sum given the information the researcher has about the M_t consumers.

Now consider what happens when the car manufacturers do not observe some portion of ω_{ijt}. In this case, if there are no other uncertainties, the researcher would model a firm's pricing decision as based on what the firm expects demand to be:

$$\text{Expected demand} = q_{jt}^e = \sum_{i=1}^{M_t} E\Big(U(x_{jt}, p_{jt}, \omega_{ijt}) \geqslant \max_{k \neq j} U(x_{kt}, p_{kt}, \omega_{ikt}; \theta)\Big). \tag{72}$$

In this expression, E is the firm's expectation over the unobservables in ω_{ijt}. (Here, the firm is assumed to know the size of the market M_t.) The firm's expected aggregate demand for model j can equivalently be expressed as the sum of firms' probability assessments that consumers will buy model j:

$$q_{jt}^e = \sum_{i=1}^{M_t} \Pr(i \text{ buys new car } j). \tag{73}$$

This expression shows us how firms' uncertainties about their environment (i.e., their uncertainties about consumers tastes) will enter a structural model of competition. In essence, the IO researcher must now take a stand on firms' beliefs about consumers – what they know and do not know – and how this information enters consumers' tastes.

Once the researcher adopts a specific probability model for consumers' product choices, product-level demands simply are sums of consumers' purchase probabilities. These purchase probabilities and sums have the potential drawback that they may be nonlinear functions of consumer taste parameters. Despite this complication, the above formulation has one important advantage. A discrete choice model allows the researcher to model consumers' preferences over a large number of products as a function of a short list of product attributes (the x_{jt}). Thus, in contrast to the high-dimensionality of the neoclassical model, here a researcher may be able to reduce the consumers' choice problem to a choice over a few attributes.

Two crucial questions that must be answered when developing a discrete choice model are: What is it about consumer tastes that firms do not observe? And: What is a sensible model of firms' expectations? These are important questions because a researcher's inferences about price–cost margins may well be sensitive to the specification

of firms' uncertainties. In what follows, we use our framework for building structural models to evaluate two early differentiated product models. Both models estimate price–cost margins for new cars sold in the United States. The first, by Goldberg (1995), uses household-level new-car purchase data to estimate household-level purchase probabilities for different new car models. She assumes that these household-level probabilities are what firms use to determine aggregate demand. The second approach we consider is by Berry, Levinsohn and Pakes (1995). They do not have household-level data. Instead, they construct their demand system from product-level price and quantity data. Like Goldberg, they too base their demand estimates on sums of individual purchase probabilities. Unlike Goldberg, they match the parameters of this sum to realized new car market shares.

7.2.1. A household-level demand model

Goldberg's model of prices and quantities in the US new car market follows the logic of a homogeneous product competition model. Her estimation strategy is divided into three steps. In the first step, Goldberg estimates household-level demand functions. In the second, the household-level demand functions are aggregated to form estimates of firms' expected demand curves. In the third step, Goldberg uses the estimated expected demand curves to calculate firms' first-order conditions under the assumption that new car manufacturers are Bertrand–Nash competitors. From these first-order conditions, she can then estimate product-level marginal cost functions and price–cost markups for each new car model. The main novelty of Goldberg's paper is that she uses consumer-level data to estimate firms' expected new car demands. The supply side of her model, which develops price–cost markup equations, follows conventional oligopoly models, but it is computationally more difficult because the demands and derivatives for all the cars sold by a manufacturer enter the price–cost margin equation for any one new car it sells.

7.2.2. Goldberg's economic model

Goldberg's economic model treats consumers as static utility maximizers. She computes firms' expected demand as above:

$$q_{jt}^{e} = \sum_{i=1}^{M_t} \Pr(i \text{ buys new car } j). \tag{74}$$

Goldberg of course does not observe firms' expectations. The initial step of her estimation procedure therefore seeks to replace $\Pr(\cdot)$ with probability estimates from a discrete choice model. The validity of this approach hinges both on how close her discrete choice probability model is to firms' assessments and how accurately she is able to approximate the sum of probability estimates.

To estimate household probabilities, Goldberg uses data from the US Bureau of Labor Statistics Consumer Expenditure Survey (CES). This survey is a stratified random

sample of approximately 4500 to 5000 US households per quarter. By pooling data for 1983 to 1987 Goldberg is able to assemble data on roughly 32,000 households purchase decisions. In her data she observes the vehicles a household purchases and the transaction price. She augments this consumer-level data with trade information about new and used car attributes.

A critical component of her expected demand model is the list of attributes that enter consumers' utility functions. While the transactions price is clearly a relevant attribute, economics provides little guidance about what other attributes might enter consumers' utilities. Goldberg's approach is to rely on numerical measures found in car buyer guides. These measures include: horsepower, fuel economy, size, and dummy variables describing options.

In estimating the expected demands faced by new car manufacturers, Goldberg relies on the representativeness and accuracy of the Consumer Expenditure Survey. The assumption that her probability model replicates the firms' assessments of consumer behavior allows her to replace $\Pr(k \text{ buys new car } j)$ in (74) with an econometric estimate, $\hat{\Pr}(k \text{ buys new car } j)$, which is sample household k's purchase probability. The assumption that the CES sample is representative of the M_t consumers allows her to replace the sum over consumers in (75) with a weighted sum of the estimated household purchase probabilities:

$$\text{Estimated expected demand for product } j = \sum_{k=1}^{S_t} w_{kt} \hat{\Pr}(k \text{ buys new car } j), \quad (75)$$

where the w_{kt} are CES sampling weights for sample household k and S_t is the number of sample households in year t.

On the production side, Goldberg assumes that new car manufacturers maximize static expected profits by choosing a wholesale price. Unfortunately Goldberg does not observe manufacturers' wholesale prices. Instead, she observes the transactions prices consumers paid dealers. In the US, new car dealers are independent of the manufacturer. The difference between the retail transaction price and the wholesale price thus reflects the independent dealer's markup on the car. The dealer's incentives are not modeled in the paper for lack of data. Because Goldberg is modeling manufacturer's pricing decisions (and not transactions prices), Goldberg assumes that there is an exact relation between the unobserved wholesale prices and average transactions prices she computes from the CES. For example, she assumes that the wholesale price of an intermediate-size car is 75% of an average transaction price she can compute from the CES. While this assumption facilitates estimation, it is unclear exactly why it is profit-maximizing for dealers and manufacturers to behave in this way.[13]

Goldberg models manufacturers' decisions about wholesale prices as outcomes of a static Bertrand–Nash pricing game in which manufacturers maximize expected US

[13] For more discussion of automobile dealer behavior see Bresnahan and Reiss (1985).

profits. The expectation in profits is taken over the demand uncertainty in each ω_{ijt}.[14] Thus, firm f maximizes

$$\max_{p_t^{wf}} \sum_{j=1}^{n_f} \left(p_{jt}^{w} - c_{jt}\right) E\left[q_{jt}\left(p^{w}\right)\right], \tag{76}$$

where $p_t^{wf} = (p_{1t}^{wf}, \ldots, p_{n_f,t}^{wf})$ is a vector of wholesale prices, n_f is the number of new car models offered by firm f and c_{jt} is the constant marginal production cost for new car model j. The first-order conditions that characterize manufacturers' wholesale pricing decisions have the form

$$p_{jt}^{wf} q_{jt}^{e} + \sum_{k=1}^{n_f} \frac{p_{kt}^{wf} - c_{kt}}{p_{kt}^{wf}} q_{kt}^{e} \epsilon_{kjt} = 0, \tag{77}$$

where $q_{kt}^{e} = E(q_{kt})$, and $\epsilon_{kjt} = \frac{p_{jt}^{wf}}{q_{kt}^{e}} \frac{\partial q_{kt}^{e}}{\partial p_{jt}^{wf}}$ is the cross-price elasticity of expected demand. This equation shows that in order to obtain accurate estimates of the firm's price–cost margins, we need to have accurate estimates of the firms' perceived cross-price elasticities. Changes in the demand model, say by changing the model of firm uncertainty about consumer tastes, will likely change the estimated cross-price elasticities, and thus in turn estimates of price–cost markups.

Once Goldberg has estimated her demand model and obtained expressions for the cross-price elasticities, the only remaining unknowns in the firms' first-order conditions are their marginal costs, the c_{jt}. Because Goldberg has one first-order condition for each product, she can in principle solve the system of equations exactly to obtain estimates of the c_{jt} and price–cost margins.

7.2.3. The stochastic model

To estimate household purchase probabilities, Goldberg employs a nested logit discrete choice model. She assumes consumers' conditional indirect utilities have the additive form

$$U_{ijt} = U(x_{jt}, p_{jt}, \bar{\omega}_{ijt}) + v_{ijt},$$

where $\bar{\omega}_{ijt}$ are observable household and product characteristics and v_{ijt} is a generalized extreme value error. Goldberg goes on to assume that the indirect utility function is linear in unknown taste parameters, and that these taste parameters weight household characteristics, vehicle attributes and interactions of the two. The generalized extreme

[14] In principle, the firm also might be uncertain about its marginal cost of production. Goldberg can allow for this possibility only if the cost uncertainty is independent of the demand uncertainty. Otherwise, Goldberg would have to account for the covariance of demand and costs in (76).

value error assumption appears to be made because it results in simple expressions for the firms' expectations about consumer purchase behavior found in Equation (74).

The generalized extreme value error results in a nested logit model. Goldberg's choice of logit nests is consistent with but does not imply a particular sequential model of household decision making. Specifically, she expresses the probability that household k buys model j as a product of conditional logit probabilities:

$$\begin{aligned}
&\Pr(k \text{ buys new car } j)\\
&\quad = \Pr(k \text{ buys a car}) \times \Pr(k \text{ buys a new car} \mid k \text{ buys a car})\\
&\qquad \times \Pr(k \text{ buys new in segment containing } j \mid k \text{ buys a new car})\\
&\qquad \times \Pr(k \text{ buys new from } j\text{'s origin and segment} \mid k \text{ buys new}\\
&\qquad\qquad \text{in segment containing } j)\\
&\qquad \times \Pr(k \text{ buys } j \mid k \text{ buys new from } j\text{'s origin and segment}). \qquad (78)
\end{aligned}$$

This particular structure parallels a decision tree in which household k first decides whether to buy a car, then to buy new versus used, then to buy a car in j's segment (e.g., compact versus intermediate size), then whether to buy from j's manufacturer – foreign or domestic, and then to buy model j.

Goldberg appears to favor the nested logit model because she is uncomfortable with the logit model's independence of irrelevant alternatives (IIA) property. The IIA property of the conventional logit model implies that if she added a car to a consumer's choice set, it would not impact the relative odds of them buying any two cars already in the choice set. Thus, the odds of a household buying a Honda Civic relative to a Toyota Tercel are unaffected by the presence or absence of the Honda Accord. The nested logit corrects this problem by limiting the IIA property to products within a nest.

In principle, Goldberg could have chosen a different stochastic distribution for consumers' unobserved tastes, such as the multivariate normal. Goldberg makes it clear that she prefers generalized extreme value errors because they allow her to use maximum likelihood methods that directly deliver purchase probability estimates. Specifically, the nested logit model permits her to compute the right-hand side probabilities in (78) sequentially using conventional multinomial logit software. Her choice of nesting structure is important here because the IIA property holds at the household level for each new car within a nest. Changes in the nests could affect her estimates of cross-price elasticities. Unfortunately, economic theory cannot guide Goldberg's nesting structure. This ambiguity motivates Goldberg to explore at length whether her results are sensitive to alternative nesting structures.

While the independence of irrelevant alternatives applies to some household choices, it does not apply at the market demand level. This is because Goldberg interacts income and price with household characteristics. By using interactions and aggregating using household sampling weights, Goldberg insures that her product-level demand functions do not have the economically unattractive IIA structure.[15]

[15] This can be seen by examining the population odds of buying two different vehicles.

Goldberg makes two other key stochastic assumptions when she estimates her nested logit model. The first is that new car prices and nonprice attributes are independent of consumers' unobserved tastes, the ν_{ijt}. This is a critical modeling assumption, as it is possible to imagine cases where it would not hold. Suppose, for instance, that the ν_{ijt} includes consumer perceptions about a car's quality, and that firms know consumers perceptions. In this case, firms' pricing decisions will depend on the car's quality. Because Goldberg does not observe quality, her econometric specification will attribute the effects of quality to price and nonprice attributes. This results in the same endogeneity problem found in neoclassical demand models. To see the parallel, imagine that ν_{ijt} consists of a product-time fixed effect ("quality") and noise. That is, $\nu_{ijt} = \xi_{jt} + \eta_{ijt}$. Because ξ_{jt} is common to all households and known to the firm, it will appear in the aggregate demand curve

$$q^e_{jt}(\xi_{jt}) = \sum_{i=1}^{M_t} \Pr(i \text{ buys new car } j \mid \xi_{jt})$$

that the manufacturer uses when choosing wholesale prices to maximize profits. Thus, wholesale prices will depend on unobserved quality. Because Goldberg does not observe product quality, she needs to devise a strategy for removing any potential correlation between price and consumers' unobserved tastes.

The best way to account for this unobserved heterogeneity within a nested logit model would be to add behavioral equations to the model that would explain how manufacturers jointly choose price and quality. Such a formulation unfortunately complicates estimation considerably. As an alternative, Goldberg could simply assume a distribution for quality and then integrate quality out of aggregate demand using this assumed distribution. This strategy is economically unattractive, however, because one would have to recognize the unknown correlation of prices and qualities when specifying the joint distribution. What Goldberg does instead is assume that unobserved quality is perfectly explained by a short list of time-invariant product characteristics, such as the manufacturer's identity (e.g., Toyota), the country of origin (e.g., Japan) and the car's segment (e.g., compact). The assumption of time invariance allows her to use fixed effects to capture these components. The ultimate question with this strategy that cannot be easily answered is: Do these fixed effects capture all the product-specific unobservables that might introduce correlation between prices and consumers' unobserved preferences? Goldberg provides arguments to suggests that these fixed effects are adequate. In principle, if she had a dataset that contained many purchases of each model, she could include a complete set of model-specific dummy variables, and thereby control for all unobserved (to the researcher) quality differences across models.

A final stochastic component of the model pertains to manufacturers' marginal costs. The system of first-order conditions (77) exactly identifies each product's marginal costs. Following Hausman, Leonard and Zona (1994), Goldberg uses these marginal cost estimates to calculate product price–cost markups, which she finds to be somewhat on the high end of those reported in other studies.

Goldberg also is interested in assessing how marginal costs are related to vehicle characteristics and market conditions. To do this, she assumes that the implied marginal costs which she recovers depend on observable product characteristics and an unobservable according to

$$c_{jt} = c_0 + Z_{jt}\alpha + u_{jt},$$

where the Z_{jt} are observable product characteristics and u_{jt} are unobservable factors affecting costs. The error in this relation accounts for the fact that the estimated marginal costs are not perfectly explained by observables.

7.2.4. Results

If we compare Goldberg's model to homogeneous product competition and neoclassical differentiated product models, we see that Goldberg's competition model is considerably richer. Her demand system (75) admits complicated substitution patterns among products. These substitution patterns depend on the proximity of products' attributes. There are two main costs to this richness. First, she must introduce many functional form and stochastic assumptions to limit the scale and computational complexity of the model. As we argued earlier, structural modelers often must introduce assumptions to obtain results. Without these assumptions and restrictions, it would be impossible for Goldberg to estimate demand and costs, or evaluate the impact of the voluntary export restraints. She also might not be able to argue convincingly that her estimates make sense (e.g., that they imply a pure-strategy equilibrium exists or is unique).

A second cost of the richness of her model is that it becomes difficult for her to summarize exactly how each economic and stochastic assumption impacts her conclusions. For example, at the household level she maintains IIA within nests. Her utility specifications and method of aggregation, however, imply that IIA will not hold at the aggregate level. But just how much flexibility is there to the aggregate demand system and the cross-price elasticities? Questions about the role of structural assumptions such as this are very difficult to answer in complex models such as this. For this reason, Goldberg and other structural modelers must rely on sensitivity analyses to understand how their conclusions depend on their assumptions. For instance, Goldberg spends considerable time exploring whether her parameter estimates and implied markups agree with other industry sources and whether the estimates are sensitive to alternative plausible structural assumptions.

While structural researchers can in many cases evaluate the sensitivity of their estimates to specific modeling assumptions, some aspects of structure are not so easily evaluated. For example, Goldberg's model relies on the maintained assumption that the weighted sum of estimated CES sample purchase probabilities accurately measures firms' expectations about product demand. If there is something systematic about firms' expectations that her household model does not capture, then this will mean she is not solving the same first-order profit maximization problems that the firms were when

they set prices. Her reliance on this assumption is nothing new. The correct specification of demand is implicit in other papers in this area (e.g., Porter and Hausman). As we argued earlier in laying out our framework, all structural models base their inferences on functional form and stochastic assumptions that are in principle untestable. In this case, Goldberg's problem is that she does not observe firms' expectations. Consequently, when she finds that her model under-predicts total new car sales, she cannot know for sure whether this is because firms underpredicted demand or there is a problem with her econometric specification or data.[16]

7.3. A product-level demand model

Bresnahan (1981, 1987) was the first IO economist to use the discrete-choice demand model in an oligopoly equilibrium model to estimate the extent of firm market power. Bresnahan's preferences assume that consumers trade off the price of the product against a single unobserved quality index. Berry (1994) and Berry, Levinsohn and Pakes (1995) (BLP) extended Bresnahan's single-index model to allow products to be horizontally differentiated. In what follows, we describe BLP's (1995) original model and compare it to Goldberg's model and the neoclassical demand systems discussed earlier. Unlike Goldberg, Bresnahan and BLP only have access to product-level data. Specifically, they know a new car model's: unit sales, list price, and attributes. BLP, for example, have twenty years of data covering 2217 new car models. Their definition of a new car model (e.g., Ford Taurus) is rich enough to describe important dimensions along which new cars differ. Their data, however, do not capture all dimensions, such as difference in some two-door versus four-door models, and standard versus "loaded" models.

BLP use these product-level price and quantity data to draw inferences about consumer behavior and automobile manufacturers' margins. Like Goldberg, they base their demand system on a discrete choice model of consumer choices. At first this may seem odd – how can they estimate a consumer choice model with aggregate data? The answer lies in the structural assumptions that permit them to relate household decisions to product-level price and quantity data.

We can informally contrast Goldberg and BLP's approaches by comparing how they model the product demands on which firms base their pricing decisions. Recall Goldberg computes firms' expected product demands as follows:

$$q_{jt}^{e} = \sum_{i=1}^{M_t} \Pr(i \text{ buys new car } j) = \sum_{i=1}^{M_t} \Psi(p_{0t}, \ldots, p_{Jt}, x_{0t}, \ldots, x_{Jt}, \bar{\omega}_{ijt}; \theta), \quad (79)$$

where the $\Psi(P, x, \bar{\omega}_{ij}; \theta)$ are the nested logit purchase probabilities that depend on the price, p, and nonprice attributes, x, of all models. Because Goldberg only uses

[16] Goldberg's chief hypothesis is that the household CES data under-represent total sales because they do not include government, business or other institutional sales.

household-level data, there is no guarantee that when she aggregates her probability estimates to form q^e_{jt} that they will match actual aggregate US sales figures, q_{jt}.

BLP (1995) on the other hand do not have the household-level data required to estimate how household choice probabilities vary with $\bar{\omega}_{ijt}$. Instead, they treat actual sales, q_{jt}, as though it is a realization from the demand curve that the firm uses to set price. In essence, they assume $q_{jt} = q^e_{jt}$. BLP then replace the household-specific probabilities $\Pr(P, x, \bar{\omega}_{ij}; \theta)$ on the right-hand side with unconditional purchase probabilities $\mathcal{S}_j(P, x, \theta)$. They do this by assuming a distribution, $P(\bar{\omega}_{ijt}, \gamma)$, for the household variables that they do not observe. Here γ denotes a set of parameters that indexes the density. Formally, they compute the unconditional demand functions

$$q^e_{jt} = \sum_{i=1}^{M_t} \int_{\omega} \Phi(p_t, x_t, \omega; \theta)\, \mathrm{d}P(\omega; \gamma) = M_t \mathcal{S}_j(p_t, x_t; \theta, \gamma), \tag{80}$$

where $\Phi(\cdot)$ are choice probabilities. The second equality follows because by assumption the distribution of consumer variables is the same for each of the M_t households in the market for a new car. To estimate the demand parameter vector θ and distribution parameter vector γ, BLP match the model's predicted expected sales $q^e_{jt} = M_t \mathcal{S}_j$ to observed sales q_{it}. (This is the same as matching expected product shares $\mathcal{S}_j$ to realized product market shares, q_{it}/M_t.) As in Goldberg's econometric model, the economic and stochastic assumptions that go into the construction of $\Pr(\cdot)$ and $\mathcal{S}_j$ have a critical bearing on the resulting demand and markup estimates.

It is useful to reiterate the differences in data requirements and modeling assumptions between Goldberg and BLP. BLP fit their model to match aggregate market shares, where market shares are national sales divided by a hypothesized number of potential buyers at time t, M_t. Consequently, the reliability of demand estimates obtained will depend on the quality of the estimates of M_t. This in turn will impact the reliability of their cost estimates. In contrast, Goldberg fits a household-level model and does not require market-level data. But as noted earlier, this data set excludes some purchases by businesses and government agencies that are relevant to firms' pricing decisions. This could impact her cost estimates.

7.3.1. The economic model in BLP

BLP's economic model of automobile sales maintains that manufacturers sell new cars directly to consumers. Manufacturers do not price discriminate and consumers are assumed to know the prices and attributes of all new cars. There are no intertemporal considerations for either firms or consumers. In particular, there is no model of how firms choose product attributes, and consumers do not trade off prices and product attributes today with those in the future.

As before, consumer i's conditional indirect utility function for new cars has the form

$$U_{ijt} = U(x_{jt}, p_{jt}, \omega_{ijt}).$$

Consumers decide to buy at most one new car per household. There are no corporate, government or institutional sales. In contrast to Goldberg, BLP do not model the choice to buy a new versus a used car. Instead, purchases of used vehicles are grouped with the decision to purchase a hypothetical composite outside good labeled product 0. Consumers demand the outside good if they do not buy a new car. Thus, if $\sum_{j=1}^{J} q_{jt}$ is the observed number of new cars bought in year t, $q_{0t} = M_t - \sum_{j=1}^{J} q_{jt}$ is the number choosing to purchase the outside good.

The firm side of the market in BLP is similarly straightforward. Sellers know the demand functions calculated above and each others' constant marginal costs of production. Sellers maximize static profit functions by choosing the price of each model they produce. When choosing price, sellers act as Bertrand–Nash competitors, as in Goldberg.

7.3.2. The stochastic model

There are three key sets of unknowns in BLP's model: the number of consumers in each year, M_t; the distribution of consumer characteristics $\Pr(\omega; \gamma)$; and sellers' manufacturing costs. We consider each in turn.

Not knowing M_t, the overall size of the market, is a potential problem because it relates the choice probabilities described in Equation (80) to unit sales. BLP could either estimate M_t as part of their econometric model or base estimation on some observable proxy for M_t. Although the first of these approaches has reportedly been tried, few if any studies have had much success in estimating the overall size of the market. This difficulty should not be too surprising because the absence of data on the outside good means that additional assumptions will have to be introduced to identify the overall size of the market.

One way to develop intuition for the assumptions needed to estimate M_t in a general model is to consider the role M_t plays in a cross-section logit model. Specifically, suppose that utility consists of an unobserved product attribute ξ_j and an extreme value error η_{ij}:

$$U_{ij} = \xi_j + \eta_{ij}. \tag{81}$$

To obtain the unconditional purchase probabilities $\mathcal{S}_j(p, x; \theta, \delta)$ we integrate out the consumer-level unobservables

$$\mathcal{S}_j = \int_{-\infty}^{\infty} \prod_{k \neq j} F(\xi_j - \xi_k + \tau) f(\tau) \, \mathrm{d}\tau, \tag{82}$$

where $F(\xi_j - \xi_k + \tau) = \Pr(\xi_j - \xi_k + \tau > \eta_{ik})$ and $f(\cdot)$ is the density of η. The integral in (82) yields the logit probabilities

$$\mathcal{S}_j = \frac{\exp(\xi_j)}{\sum_{k=0}^{J} \exp(\xi_k)}. \tag{83}$$

The demand functions are then

$$q_j = M\mathcal{S}_j(\xi_0, \ldots, \xi_J) \tag{84}$$

or using (83)

$$\ln q_j = \ln M + \xi_j - \ln\left(\sum_{k=0}^{J} \exp(\xi_k)\right). \tag{85}$$

The demand parameters here are $\theta = (\xi_0, \xi_1, \ldots, \xi_J, M)$. As a simple counting exercise, we have J equations in J observed new vehicle quantities, and $J + 2$ unknowns, $\theta = (\xi_0, \xi_1, \ldots, \xi_J, M)$. Adding a quantity equation for the unobserved quantity of the outside good, q_0, does not change the difference between unknowns and knowns, but does allow us to collapse the log-quantity equations to

$$\ln q_j - \ln q_0 = \xi_j - \xi_0. \tag{86}$$

Since by definition $q_0 = M - \sum_{j=1}^{J} q_j$, we can rewrite the J equations as

$$\ln q_j - \ln\left(M - \sum_{j=1}^{J} q_j\right) = \xi_j - \xi_0. \tag{87}$$

In general, we require at least two restrictions on the $J+2$ unknown demand parameters $(\xi_0, \xi_1, \ldots, \xi_J, M)$ to be able to solve these J equations. Since the outside good is not observed, we can without loss of generality normalize ξ_0 to zero. This still leaves us one normalization short if M is unknown.

In their empirical work, BLP choose to fix M rather than restrict the ξ's or other parameters. Specifically, BLP assume that M_t is the total number of US households in year t. This choice has some potential shortcomings. Not all households can afford a new car. As in Goldberg, entities other than households purchase new vehicles. In principle, one could model these discrepancies by assuming that the total number of US households is a noisy measure of M_t, i.e., $\widetilde{M}_t = M_t + \Delta_t$. Substituting $\widetilde{M}_t$ into (87) with $\xi_0 = 0$ gives

$$\ln q_j - \ln\left(\widetilde{M}_t - \sum_{j=1}^{J} q_j\right) = \tilde{\xi}_j. \tag{88}$$

If we overestimate the size of the market (i.e., $\widetilde{M}_t > M_t$) then the left-hand side is smaller than it would otherwise be by the same amount for each product. This will make the average (unobserved) ξ_j seem lower, or in other words that all new cars that year are worse than average. In essence, the unobserved product qualities would act as a residual and capture both true quality differences and measurement error in the size of the market.

In actual applications, we will never know whether we have over-estimated or under-estimated M_t. This means that we will not know the direction of any bias in estimated

product qualities, the ξ_j's. While the availability of panel data might allow us to attempt a random measurement error model for M_t, in practice the nonlinearity of the demand functions in the measurement error will make it difficult to draw precise conclusions about how this measurement error impacts demand estimates. Thus, one is left with either using a proxy for M_t as though it had no error or imposing enough additional restrictions on the demand model so that M_t can be estimated.

The second set of unobservables that enter BLP's demand functions are the household variables, ω_{ijt}. Formally, BLP assume household i's indirect utility for new car j has the additive two-part structure:

$$\begin{aligned} U_{ijt} &= \overbrace{\qquad\delta_{jt}\qquad}^{} + \overbrace{\qquad\qquad\omega_{ijt}\qquad\qquad}^{} \\ &= \overbrace{x_{jt}\beta + \xi_{jt}}^{} + \overbrace{x_{jt}\nu_i + \alpha \ln(\nu_{iy} - p_{jt}) + \eta_{ijt}}^{}. \end{aligned} \tag{89}$$

The δ_{jt} includes only product attributes. For BLP it consists of a linear function of observed (x) and unobserved (ξ) product attributes. The elements of the $K \times 1$ parameter vector β are interpreted as population average marginal utilities for the observed attributes.

The ω_{ijt} contain three separate household-level terms. The familiar extreme value error term η_{ijt} allows for unobserved household-specific tastes for each model in each year. The $K \times 1$ vector of unobservables ν_i allows for the possibility that household i's marginal utilities for observed attributes differ from the population average marginal utilities (the β's). While in principle one might expect that households' marginal utilities would depend on household income and other demographic characteristics, the lack of household data forces BLP to assume that the ν_i's are independent random variables that are identically distributed in the population.[17] BLP assume these random variables are normally distributed. In addition, they assume that a household's unobserved marginal utility for attribute k is independent of their marginal utility for attribute h. The unboundedness of the support of the normal distribution implies that some households will prefer attribute k and some will have an aversion to it. Specifically, the fraction that dislike attribute k is given by $\Phi(-\beta_k/\sigma_{ik})$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function and σ_{ik} is the standard deviation of ν_{ik}.

The final stochastic component of ω is $\alpha \ln(\nu_{yi} - p_{jt})$, where α is an unknown coefficient. We use the notation ν_{yi} for income to indicate that, like the unobserved marginal utilities for the observed attributes, income also is an unobservable. The expression in the natural logarithm is the (unobserved) income the household has left if it purchases model j. BLP include $\ln(\nu_{yi} - p_{jt})$ so that they can interpret $U_{ijt}(\cdot)$ as a conditional indirect utility function. Once again they need to make some distributional assumption on the unobserved ν_{yi} in order to compute expected demand. In their empirical work BLP assume that the natural logarithm has a lognormal distribution. However, the lognormal distribution must be truncated to make the expenditure on the outside good

[17] BLP and others have explored alternatives to this structure. For example, BLP (2004) allow consumers' marginal utilities to depend on observable and unobservable household attributes.

positive. That is, they need to guarantee that $v_{yi} > p_{jt}$ for all observed and plausible counterfactual prices p_{jt}.

A final element of the preference specification is BLP's treatment of the outside good. BLP assume that the utility for the outside good has the form:

$$U_{i0t} = \alpha \ln v_{iy} + \sigma_0 \nu_{i0} + \eta_{i0t}.$$

Unobserved income enters this utility because it is the amount available to be spent on the outside good when no new car is purchased. No price enters the conditional indirect utility for the outside good because p_0 has been assumed to equal zero. The parameter σ_0 is new; it represents the standard deviation of the household's unobserved preference for the outside good, ν_{i0}. In essence, ν_{i0} increases or decreases the unobserved product qualities, the ξ_j, for household i by the same amount. By adding the same household-specific constant to the ξ's, BLP preserve households' rankings of all new cars based on their unobserved qualities, but allow households to disagree on the overall quality of new cars. To see this, suppose for simplicity that $\alpha = 0$ and $\beta = \nu_i = 0$. Utilities then are as in Equation (89) except $U_{i0} = \sigma_0 \nu_{i0} + \eta_{i0}$. The logit probabilities of purchase in (83) now have the household-specific form

$$\mathcal{S}_{ij} = \frac{\exp(\xi_j - \sigma_0 \nu_{i0})}{1 + \sum_{k=1}^{J} \exp(\xi_k - \sigma_0 \nu_{i0})}. \tag{90}$$

Thus, households with large values of ν_{0i} do not think that the quality of new cars is very high and consequently are more likely to opt for the outside good. Similarly, holding the unobserved qualities of new cars fixed (the ξ), increases in σ_0 reduce the importance of the unobserved car model qualities for purchase decisions.

7.4. More on the econometric assumptions

Now that we have provided an overview of BLP's economic and stochastic assumptions, it is useful to revisit some of them to understand further why BLP adopt these assumptions.

7.4.1. Functional form assumptions for price

A critical component of any choice model is the way in which product prices enter utility. Consider what would happen, for example, if BLP had entered (as some studies do) price as an additive function in δ_{jt} rather than in ω_{ijt}. In a standard logit choice model, with $\delta_{jt} = g(p_{jt}) + \tilde{\delta}_{jt}$, the demand equations have the form

$$\ln q_{jt} = \ln M_t + g(p_{jt}) + \tilde{\delta}_{jt} - \ln\left(1 + \sum_{k=1}^{J} \exp\bigl(g(p_{kt}) + \tilde{\delta}_{kt}\bigr)\right). \tag{91}$$

The implied own-price and cross-price elasticities for these demands are:

$$\frac{\partial \ln q_{jt}}{\partial \ln p_{kt}} = \begin{cases} \frac{\partial g(p_{jt})}{\partial p_{jt}} p_{jt}(1 - \mathcal{S}_{jt}), & k = j, \\ -\frac{\partial g(p_{jt})}{\partial p_{kt}} p_{kt}\mathcal{S}_{kt}, & k \neq j. \end{cases} \tag{92}$$

These expressions show how the extreme value error assumption and the choice of $g(\cdot)$ affect the structure of the own-price and cross-price elasticities that enter the price-markup equations. If price enters logarithmically (e.g., $g(p_{jt}) = \theta \ln p_{jt}$), then the own-price and cross-price elasticities only depend on product market shares. In this case, an increase in the price of a Jaguar would cause the demand for BMWs and Kias, which have roughly similar shares, to increase roughly the same amount, even though BMWs and Kias are hardly substitutes. To some extent, one could consider fixing this problem by changing the way price enters δ_{jt} or by interacting functions of price with other vehicle attributes. Such an approach, however, ultimately may not capture what one might expect, which is that products with similar attributes will have higher cross-price elasticities.

The use of the extreme value error can also have some other unattractive economic consequences. One consequence of the error's unbounded support is that with finite attributes, there always will be someone who will buy a product – no matter how inferior the car is to other cars. Suppose, for example, that instead of having price enter logarithmically, the function $g(p)$ is bounded above. In this case, product demands will asymptote to zero instead of intersecting the price axis. This asymptotic behavior can have an unfortunate impact on global welfare and counterfactual calculations. Petrin (2002), for example, finds that when price is entered linearly that one can obtain implausibly large estimates of the value of Minivans. Figure 1 illustrates this problem for two alternative specifications of $g(\cdot)$ using a standard logit model for shares. The demand curve labeled A assumes price enters δ as $-\lambda p$. The concave demand curve B adopts a logarithmic specification paralleling BLP, $g(p) = \lambda \ln(100 - p)$. The constant λ is selected so that each curve predicts roughly the same demand for a range of prices between 60 and 90. (One might think of this as approximating a range of data that the researcher would use to estimate λ.) Comparing the two demand curves, we can see that there would not be too much of a difference in the two models' predicted demands or local consumer surplus calculations for prices between 60 and 90. But often researchers perform calculations that rely on the shape of the demand function for all positive prices. A common example is determining the welfare gain to consumers from the introduction of a new good [e.g., Petrin (2002) and Hausman (1997)]. In this case, the properties of the demand function for the new good from the price where the demand curve intersects the vertical axis to the actual market price determine the benefits consumers derive from the existence of this good. The difference in this calculation for the two demand curves would be dramatic. For example, Demand Curve A estimates that there are many consumers with reservation prices above 100, while Demand Curve B says there are none.

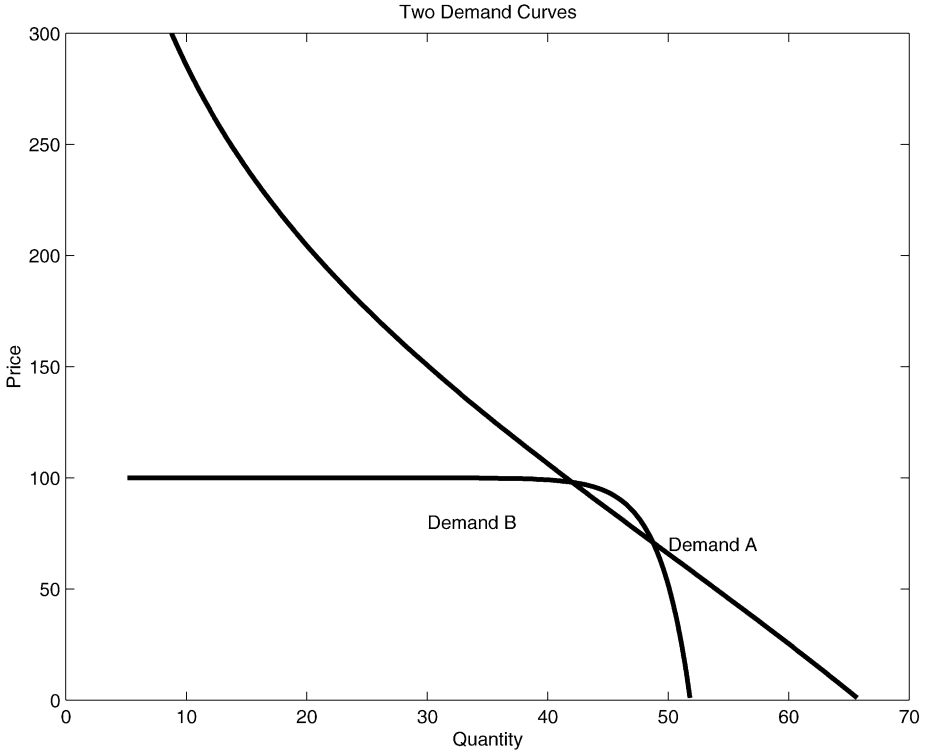

Figure 1.

7.4.2. *Distribution of consumer heterogeneity*

In their empirical work, BLP emphasize that they are uncomfortable with the IIA property of the standard logit choice model, and for this reason they add unobservable household-car attribute interactions. To gain some understanding of what these unobservables add, consider the following three good market:

- there are two types of cars available: large (LARGE $= 2$) and small (LARG$E = 1$);
- utilities for the large and small cars equal

$$U_{ij} = \beta_0 + \beta_L \text{LARGE}_j + \eta_{ij} = \delta_j + \eta_{ij};$$

  and
- the large car has 15 percent of the market, the small car 5 percent and the outside good the remaining 80 percent.

This utility specification perfectly explains the market shares. That is, we can match the observed market shares to the logit shares exactly:

$$0.15 = \exp(\beta_0 + 2\beta_L)/\big(1 + \exp(\beta_0 + \beta_L) + \exp(\beta_0 + 2\beta_L)\big),$$
$$0.05 = \exp(\beta_0 + \beta_L)/\big(1 + \exp(\beta_0 + \beta_L) + \exp(\beta_0 + 2\beta_L)\big). \tag{93}$$

A solution is: $\beta_L = 1.098$, and setting $\beta_0 = -3.871$. Although the deterministic utility specification predicts consumers prefer larger to smaller cars, the infinite support of the extreme value error ν_{ijt} results in some consumers having an idiosyncratic preference for small cars.

Now consider what happens with these data when we add heterogeneity in consumers' marginal utilities for size. In lieu of assuming a continuous distribution of marginal utilities, suppose for simplicity that there are just two types of consumers: those with a taste β_{L1} for size and those with a taste β_{L2} for size. Because we can potentially explain the three market shares with just two parameters, assume $\beta_0 = 0$. In addition, to avoid the complication of having to estimate the entire distribution of consumer preferences, suppose we know that 15 percent of consumers are of type 1 and the remaining 85 percent are type 2.

How does this two-type model explain the market share of the small car? It seems in principle that the two-type model could fit the market share data in the same way that the single type model did. Both types of consumers would have positive but different marginal utilities for vehicle size, and the unbounded support of the extreme value error would account for why some fraction of each type would buy an otherwise inferior car. To see whether this is the case, we again match the observed market shares to the logit shares:

$$0.15 = 0.15\frac{\exp(2\beta_{L1})}{1 + \exp(\beta_{L1}) + \exp(2\beta_{L1})} + 0.85\frac{\exp(2\beta_{L2})}{1 + \exp(\beta_{L2}) + \exp(2\beta_{L2})},$$
$$0.05 = 0.15\frac{\exp(\beta_{L1})}{1 + \exp(\beta_{L1}) + \exp(2\beta_{L1})} + 0.85\frac{\exp(\beta_{L2})}{1 + \exp(\beta_{L2}) + \exp(2\beta_{L2})}. \tag{94}$$

A solution is the type 1 consumers have a negative marginal utility for size ($\beta_2 = -2.829$) and the type 2 consumers have a positive marginal utility for size ($\beta_1 = 3.9836$). Thus, when consumers' marginal utilities are unconstrained, the choice model may explain the purchase of an inferior product by indicating that some consumers have negative marginal utilities for otherwise attractive attributes.

This example gets at the heart of IO economists' distinction between vertical and horizontal product differentiation models. In vertical models, consumers share similar opinions about an attribute, and thus will rank products the same. They may, however, differ in the strength of their preferences. In multi-attribute models, the relation between vertical and horizontal product differences and product rankings becomes more complex. For instance, even though consumers may all have positive marginal utilities for all attributes, they may rank products differently.

In most applications researchers will have only a few attributes that they can use to explain why consumers prefer one product over others. When there are many products

compared to attributes, a large number of products may appear "dominated" according to a pure vertical model. For example, the Volkswagen Beetle is a small car, has a small engine, slightly higher than average fuel economy, etc., and yet at times sold relatively well in the US. One way BLP's model could explain the apparent relative success of the Beetle would be to assign it a high unobserved quality, ξ. Alternatively, as we have seen above, the introduction of heterogeneous tastes can account for why consumers might prefer an otherwise "average" or "dominated" product. While the introduction of consumer heterogeneity can increase the flexibility of a discrete choice model, this increased flexibility may or may not lead to results that are economically plausible. For instance, in BLP's Table IV (p. 876), they report an estimated distribution of marginal utility for miles per dollar (MP$) across consumers that is normal with mean -0.122 and a standard deviation 1.05. This estimate implies that roughly 54 percent of consumers "dislike" fuel economy, in the sense of having a negative marginal utility of miles per dollar. For the marginal utility of horsepower divided by weight of the car (HP/Weight), the estimated distribution of marginal utility is normal with mean 2.883 and standard deviation 4.628. This implies that 27 percent of consumers dislike cars with higher values of HP/Weight. Using BLP's assumption that these marginal utilities are independent implies that 14 percent of consumers prefer cars with lower values of HP/Weight and higher values of MP$. The plausibility of these estimates of the distribution customer-level heterogeneity is an open question. However, it is important to bear in mind that the assumptions BLP make about the functional form of customer's demands, the joint distribution of customer marginal utilities and income identify the joint distribution of marginal utilities that BLP recover. In contrast, one advantage of Goldberg's approach which uses household-level data, is that the marginal utilities can be identified from the household-level purchases probabilities that she estimates.

Because inferences about consumer heterogeneity are conditional on maintained functional form assumptions, it seems imperative that some effort should go into exploring the robustness of findings to distributional assumptions. To date, there has been only a modest amount of effort along these lines [see Ackerberg and Rysman (2005), Berry (2001), Bajari and Benkard (2001, 2005) and the references therein], and much more work remains to be done. In their empirical work, BLP appear to prefer the use of normal distributions because it simplifies computations. However, their computations appear to be simplified more by their assumption that marginal utilities are independent, than their assumption of normality.

7.4.3. Unobserved "product quality"

The unobserved car attributes, the ξ_{jt}, are critical stochastic components of BLP's random utility model. Although the literature sometimes refers to the ξ_{jt} as unobserved quality, they can be any combination of product-specific unobservables that enter consumers' utility in the same way. The relevance of the ξ_{jt} is perhaps best understood by returning to the cross-section logit model where $\delta_j = \xi_j$ and $\xi_0 = 0$. In this case,

demands have the form

$$\ln q_j - \ln\left(M - \sum_{j=1}^{J} q_j\right) = \xi_j. \tag{95}$$

From this equation we see that the ξ_j act as demand "errors" that insure that the econometric choice model's predicted market shares match the observed market shares. Goldberg accounts for the presence of these ξ_j through her market segment, country-of-origin, and brand fixed effects. In BLP's model it is essential that the predicted and observed market shares match. This is because BLP's theoretical model presumes that (unconditionally) *each* consumer's decision can be represented by the same multinomial choice probabilities: $(\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_J)$. Thus, with a sample size of approximately 100 million, there should be no appreciable difference between their model's predictions and observed market shares. The only way to guarantee that there will be no difference is to have a sufficiently rich parameterization of demand. The ξ's achieve just this.

As errors, the ξ are subject to arbitrary normalizations. To understand better why normalizations are necessary, let us return to the cross section logit model. Assume that $\delta_j = x_j\beta + \xi_j$, where x_j is a $K \times 1$ vector of product attributes. Now, the J equations in (87) become

$$\ln q_j - \ln\left(M - \sum_{j=1}^{J} q_j\right) = x_j\beta + \xi_j. \tag{96}$$

Assuming M is known, we have J linear equations in $J+K$ unknowns: $(\xi_1, \ldots, \xi_J, \beta)$. We therefore require K linearly independent restrictions in order to estimate the marginal utility parameters uniquely. One choice would be to set K of the ξ's to zero. BLP instead opt to place moment restrictions on the distribution of the ξ.[18] Although they do not motivate their restrictions in any detail, the computational rationale for the restrictions is readily apparent. Specifically, BLP assume that the ξ are mean independent of the observed characteristics of new cars: $E(\xi_j \mid x_1, \ldots, x_J) = 0$. This moment condition is useful because it mimics the usual least squares moment conditions, and thus, if valid, could be used to estimate the marginal utilities (the β's) in (96). In least squares, the population moment conditions are replaced by K sample moment conditions.

While imposing the population moment condition $E(\xi_j \mid x_1, \ldots, x_J) = 0$ has a useful computational rationale, it also has nontrivial economic implications. In particular, if we view ξ as an unobserved product attribute such as product quality, then we have to wonder why it would not be correlated with observable attributes. While we can think of some attributes that might be uncorrelated, such as the number of doors on a car,

[18] In principle, BLP also could have considered other restrictions on the distribution of the ξ. For example, BLP could integrate out the population market share conditions over a distribution for the ξ_j. Such an approach is problematic when the ξ_j are correlated with observables such as price because the supply side of their model suggests a complex equilibrium relationship between price and the ξ_j.

if x_j were to include the new car's price, then there would be a clear cause for concern. The concern is one of unobserved heterogeneity – the firms observe the quality that consumers assign to cars and use this information to set price. (Intuitively, firms will set higher prices for cars with higher quality.)

BLP explicitly recognize this problem and do not include price in the list of conditioning variables $x_1, \ldots, x_J$. This means that they must introduce at least one other moment condition to estimate the price coefficient. As in the Marshallian demand case, BLP in principle have many candidate variables they can use to form moment conditions, including the attributes of other vehicles. These other attributes effectively act as "instruments" for price and any other endogenous attributes.[19]

Another question that arises in this type of study is: What guarantees that nonprice attributes are valid as instruments? This is the same issue that arose in our discussion of neoclassical demand systems. One might well imagine that car manufacturers choose attributes, such as air conditioning and size, in concert with a new car's quality (or other unobservable characteristics). If this is the case, then these attributes are no longer valid instruments. To obtain valid instruments, we would presumably need to model the determinants of product attributes.

In their empirical work, BLP base estimation on sample moment conditions involving the demand and marginal cost errors (discussed below). As can be seen from the market share expressions in Equation (80), in general it is not possible to compute closed form expressions for the δ_{jt} and ξ_{jt} that enter the population moment conditions. This means in practice that the researcher must numerically invert Equation (80) or use a fixed point algorithm to solve for the ξ_{jt}. While the integral in (80) is straightforward conceptually, it is difficult to compute in practice. As an alternative, BLP use Monte Carlo simulation methods to approximate the right-hand side integral. Specifically, they use importance sampling methods to estimate the integral in (80). They then recover the δ_{jt} using a fixed-point algorithm. From estimates of the δ_{jt}, the ξ_{jt} can be recovered from the residuals of an instrumental variable regression of δ_{jt} on product attributes.

7.4.4. Cost function specifications

To this point, we have said little about the cost side. In principle, one could estimate the demand parameters without using information from the supply side. BLP appear to add the supply side for at least two reasons. First, it contributes variables that can be used in the orthogonality conditions that identify the demand parameters. Specifically, their cost-side model contributes two additional instruments (a time trend and miles per gallon). Following the approach discussed above for constructing demand error instruments, BLP now have 21 (seven instruments times 3) sample moment conditions

[19] For example in the cross section logit model we can replace the moment condition $E(\xi_j \mid p_j) = 0$ with $E(\xi_j \mid x_{k1}) = 0$, where x_{k1} is an exogenous characteristic of car k. This again gives us K moment equations. The resulting estimator is indirect least squares, in which x_{k1} serves as an instrument for price.

for the cost-side error.[20] Second, by having a supply side model, they can study how manufacturers' marginal costs seem to vary with a model's attributes.

The stochastic specification of the cost-side is fairly straightforward. Sellers equate the marginal revenues for each model with the constant marginal costs of producing that model. The researcher estimates sellers' marginal revenues by differentiating the market share functions. As in other oligopoly models, BLP decompose product marginal cost into an observable and an unobservable component. Specifically, they assume that the natural logarithm of marginal costs depends linearly on a set of cost variables and an additive error. This error is also used to form moment conditions under the assumption that its mean does not depend on new car attributes or cost variables.

7.5. Summary

BLP report estimates for several demand models. They provide elasticity and markup estimates for different new car models. They argue that these estimates roughly accord with intuition. They also make a case for their unobserved heterogeneity specification. Because of the complexity of their model, it is harder for the authors to provide a sense for how their various maintained assumptions impact their results. For instance, the markups are predicated on the Bertrand–Nash assumption, the choice of instruments, the attribute exogeneity restrictions, the stationarity and commonality of unobserved product attributes. Subsequent work, including work by BLP individually and jointly has sought to relax some of these restrictions.[21] Ongoing work by others is exploring the consequences of other assumptions in these models, and we leave it to others to survey this work.[22]

In concluding this section on differentiated product demand estimation, we want to come back to some of the themes of our structural estimation framework. Previously we emphasized that researchers should evaluate structural models in part by how well the economic and statistical assumptions match the economic environment being studied. Differentiated product models pose an interesting challenge in this regard, both because they are difficult to formulate and because data limitations often limit the flexibility that one can allow in any particular modeling format. At present, there are few standards, other than crude sanity checks, that researchers can use to compare the wide array of assumptions and estimation procedures in use. For example, to date researchers have used both neoclassical demand and discrete choice models to estimate price elasticities and markups for ready-to-eat cereal products. Ready-to-eat cereal products would hardly seem to fit the single purchase assumption of current discrete choice models. Neoclassical models suffer from their reliance in representative-agent formulations. There also

[20] Because of near collinearity concerns, they drop two of these moment conditions in estimation. That is, they base estimation on the 5 times 3 (= 15) demand instruments plus 2 times 3 (= 6) cost instruments less two demand-side instruments.

[21] For example, Berry (2001) and Berry, Levinsohn and Pakes (2004).

[22] For example, Ackerberg and Rysman (2005), Bajari and Benkard (2001), and Bajari and Benkard (2005).

have been few attempts to date made to investigate the finite sample or asymptotic performance of different estimation procedures.[23]

8. Games with incomplete information: Auctions

Over the last thirty years, economic theorists have explored a variety of game-theoretic models in which private or asymmetric information impacts economic behavior. Examples include adverse selection, contracting and auction models. In these models, agents have private information about their "type" (e.g., productivity, health status, or valuation) and general information about the joint distribution of other agents' types. Agents may also face general uncertainty about their market environment (e.g., uncertainty over prices or aggregate productivity). Within this environment, agents use their private information strategically. The econometrician typically does not know agents' private information, market uncertainties, or the distribution of agents' private information. Thus, structural models of privately-informed decisions must take into account not only unobserved private information, but also how agents' actions are influenced by private information.

Our goal here is to illustrate how the framework in Section 4 can be used to compare different econometric models of privately informed agents. Much of this discussion focuses on auction models. Auctions have recently received enormous attention in IO, and continue to serve as a proving ground for empirical models of privately informed agents. We next discuss empirical models of regulated firms' decisions and regulator behavior. These models share similarities with auction models, but also pose special modeling issues.

8.1. Auctions overview

Our discussion of inter-firm competition models emphasized that it is economic theory that allows one to move from estimates of the conditional joint density of prices and quantities, $f(P, Q \mid X, Z)$, to statements about firms' demands, firms' costs and competition. This same principle applies to models with privately informed agents – absent economic assumptions, nothing can be said about agents' behavior or their private information.

In auction studies, economists usually know:

1. each auction's format;
2. the winning bid, and possibly all bids: $B = (b_1, b_2, \ldots, b_N)$;

[23] Indeed, with panel data on products, where new products are being introduced and old ones abandoned, it is unclear what would constitute a large sample argument for consistency or efficiency. See, however, Berry, Linton and Pakes (2004).

3. item-specific or auction-specific information X (e.g., number of potential bidders, reservation price, size, quality, date of auction); and
4. bidder-specific information, $Z = (z_1, z_2, \ldots, z_N)$ (e.g., bidders' identities and size).

In ideal applications, the economist has complete information on (B_i, X_i, Z_i) for a large number $(i = 1, \ldots, I)$ of related auctions. Thus, the number of bidders and potential bidders is known; there are no missing bids; and there are no X_i or Z_i that the bidders observe that the econometrician does not. Absent an economic model of the auction, the best an empiricist can do with these ideal data is recover a consistent estimate of $g(B_i \mid Z_i, X_i)$ – the conditional density of bids given bidder and auction characteristics.

The conditional density $g(B_i \mid Z_i, X_i)$ is a statistical object. Its dimensionality depends on the number and identity of bidders, and on the variables in X_i and Z_i. The dimension of $g(\cdot)$ is critical because in order to estimate $g(\cdot)$ nonparametrically and precisely, a researcher will need a large sample of similar auctions. The auctions must be similar in the sense that they have the same number of bidders and the same X and Z variables. If this is not the case, then the researcher must divide the sample so as to estimate a separate nonparametric function for each combination of bidders, and each set of X and Z variables. For this reason, and because auction data are rarely ideal, empiricists typically do not estimate $g(B_i \mid Z_i, X_i)$ nonparametrically. Instead, they use economic theory to place considerable structure on $g(B_i \mid Z_i, X_i)$.

In what follows we first describe how, with the aid of economic theory, one can recover economic objects from nonparametric estimates of $g(B_i \mid Z_i, X_i)$ (or objects that can be derived from the density). This discussion illustrates how differing combinations of economic and statistical assumptions can be used to identify economic constructs. Because in practice it may be difficult or impossible to estimate $g(B_i \mid Z_i, X_i)$ precisely, we next discuss why one may want to assume a parametric form for $g(B_i \mid Z_i, X_i)$.

We begin our discussion of auction models by observing that there is a strong similarity between the first-order conditions estimated in homogeneous-product oligopoly models (discussed in Sections 5 and 6) and the first-order conditions of auction models. The two types of models employ substantially different stochastic assumptions however. In structural oligopoly models, the stochastic structure typically comes from the first, second and fourth sources of error described in Section 4 – namely, researcher uncertainty about the economic environment, firm uncertainty about consumers, and measurement error. By contrast, in most auction models, the stochastic structure rests *exclusively* on the second source of model error – "agent uncertainty about the economic environment" – and that uncertainty affects strategic interactions. Understanding the ramifications of these different stochastic specifications is key to understanding how structural modelers go about recovering agents' unobserved private information from data.

In auctions, we shall see it is also important to distinguish between two types of "agent uncertainty". One is private information. In this case, bidders know something that directly affects their probability of winning. The bidders are uncertain, however, about the other bidders' private information. The other type of uncertainty is common

to all bidders. In this case, bidders do not know the "common" value of the auctioned item. (They also may have different, privately held opinions about the value of the item.)

Auction models differ not only in what agents know before they bid, but also according to what they assume about whether one bidder's information is useful to another bidder. In the simplest models, agents' private information is independently distributed and useless to other agents. In more general settings, nonnegative correlations or "affiliation" among private valuations may allow bidders to use other bidders' behavior to infer something about the unknown value of the item being auctioned. As we shall see, relationships among bidders' information can have an important bearing on what a researcher can recover from auction bids.

8.1.1. Descriptive models

IO economists have devoted substantial attention recently to analyzing auction bid data. Hendricks and Paarsch (1995), Laffont (1997), and Hendricks and Porter (in press) provide excellent introductions and surveys of empirical research on auctions. Prior to the early 1990s, empirical research on auctions largely used regressions and other statistical techniques to describe how bids, or bid summary statistics, varied with auction-specific and bidder-specific variables. Of particular interest was the effect that the number of bidders had on the level and dispersion of bids, as the number of bidders was seen to be related to the extent of competition.

The results of many of these descriptive studies were difficult to interpret. This was because it was often unclear how the data or methods in these studies disentangled differences in bids due to: observable or unobservable characteristics of bidders; strategic considerations; or simply differences in bidders' beliefs about other bidders' valuations. These problems were perhaps due to the generality in which some auction models were originally cast. Additionally, these theories were not originally developed to place testable restrictions on bid distributions. As auction theories were refined and extended in the late 1970s and 1980s, empiricists began to find the theory more useful for comparing bids from different auctions and evaluating bid summary statistics.

Hendricks and Porter (1988) provide a compelling example of how empirical researchers adapted these new auction models to data. Hendricks and Porter used bids from US government offshore oil and gas lease auctions to study the effect that the presence of more-informed bidders had on the distribution of bids. In their data, they identified more-informed bidders as those bidders who owned tracts adjacent to the auctioned tract. Their logic is that, because geologic formations with oil and gas often extend over large areas, exploration activities on adjacent tracts are likely to confer an informational advantage. To develop hypotheses, Hendricks and Porter devised a theoretical model of how less-informed bidders will behave in the presence of a single more-informed bidder. In their model, there is common uncertainty about the future value of the auctioned tract (say because the future price of oil and the amount of resources in the ground are unknown). Their theoretical model yields an equilibrium in

which the less-informed bidders use mixed strategies and the more-informed firm uses a pure strategy. From this model, they are able to derive properties of the distribution of the maximum bid by a less-informed bidder. They compare this distribution to an *ex ante* distribution of informed bids. The two differ in several ways, including the probability that there will be no bid or a bid at the reservation price. They also derive results for the probability that the more-informed bidder will win and the profits of more-informed and less-informed bidders.

In their empirical work, Hendricks and Porter account for several ways in which lease auctions differ from the auctions in their theoretical model. First, their model assumes the presence of one informed bidder, but in their data there can be multiple informed bidders. Second, their results are cast in terms of the distribution of the maximum bid by a less-informed bidder. Unfortunately, they do not know the exact structure of this distribution. These realities lead Hendricks and Porter to estimate a flexible parametric joint density of the maximum bid submitted by the more-informed (B_M) and maximum bid submitted by the less-informed (B_L) bidders. They use these estimated densities to examine certain predictions of their theoretical model. Mapping what they did to our notation, Hendricks and Porter cannot estimate and test restrictions on $g(B_i \mid Z_i, X_i)$, but they can estimate a joint density for two elements of B_i, the maximum bids of the more-informed, B_M, and less-informed, B_L, bidders.

Another practical reality is that the government sets reserve prices (or minimum bids) in these auctions. While Hendricks and Porter argue that the presence of reserve prices does not affect the equilibrium bid functions, as a practical matter Hendricks and Porter never observe more-informed and/or less-informed bids below the reserve price. That is, the reserve prices truncate the conditional density of $g(B_i \mid Z_i, X_i)$. This leads Hendricks and Porter to model the truncated distribution of maximum bids. Specifically, they assume that absent truncation, the joint distribution of B_M and B_L follows a bivariate lognormal distribution. To handle reserve prices, they work with scaled bids: $(y_{Mk}, y_{Lk})'$, where they assume

$$y_{ik} = \ln(B_{ik}/R_k) = \left(X_k' Z_{ik}'\right)\theta_i + \epsilon_{ik},$$

$i = (M, L)$, R_k is the reserve price for auction k, and $(\epsilon_{Mk}, \epsilon_{Lk})'$ are independent and identically distributed normal random errors. The variables in X_k and Z_{ik} contain tract and sometimes bidder-specific information for each auction.

The presence of the reserve price means that Hendricks and Porter only observe the y_{ik} when they are greater than or equal to zero. This type of truncation can readily be accounted for in a maximum likelihood setting using tobit-like models. In their empirical work, they develop a likelihood-based model for the scaled bids (y_{Mt}, y_{Lt}) that takes into account truncation and correlation in the bid unobservables $(\epsilon_{Mt}, \epsilon_{Lt})'$. With this amended model, they test which elements of X and Z enter the joint density of (y_{Mt}, y_{Lt}). They find a variety of results supporting their theoretical model. For instance, conditional on a set of auction-specific observables (X), the participation and bidding decisions of informed firms are more highly correlated with measures of *ex post* tract value.

8.1.2. Structural models

Hendricks and Porter's paper illustrates many of the important challenges that structural modelers face in trying to match theoretical auction models to data. Their paper also illustrates how features of the auction, such as reserve prices, may require the econometrician to make compromises. Before we can begin to evaluate different structural econometric models of auctions, we first describe the economic objects structural modelers seek to recover from auction data. After describing these economic primitives, we turn to describing various economic and statistical assumptions that have been used to recover them.

The primary goal of most structural econometric models of auctions is to recover estimates of:

1. bidders' utilities $U = (u_1, \ldots, u_N)$ (or the joint density $f_U(U)$ of these utilities); and
2. information about the uncertainties bidders face.

In single-unit auctions, bidders are modeled as receiving a nonzero utility from winning that depends on the price bidder j paid, P_j. Depending on the type of auction being modeled, bidders' utilities from winning may also depend on unobservables, such as the *ex post* value of the auctioned item. In order to proceed, the researcher thus must make some assumption about individual risk preferences. Most models assume bidders are risk neutral. In the risk neutral case, bidder j's utility can then be modeled as the difference between the *ex post* value for the object and the price the winner pays: $u_j = v_j - P_j$.

There are several critical things to note about bidders' utilities. First, it is the price paid that enters the utility function. Depending on the auction rules, there can be a difference between the amount bidder j bids, B_j, and the amount they pay, P_j. For example, in a second-price (Vickery) purchase auction, the winner pays the second-highest price, which is less than or equal to what they bid.[24] Second, as we mentioned earlier, there can be a difference in bidder j's *ex ante* and *ex post* private assessment of the value of the item. When there is no difference between bidder j's *ex ante* and *ex post* private assessment of the value of the item, we have a private values (PV) model. In this case, the v_j and their joint density, $f(v_1, \ldots, v_N) = f(V)$, are direct objects of interest. When there is a difference between bidder j's *ex ante* and *ex post* private assessment of the value of the item, this is modeled as being due to "common values", v. These common values are unobserved by the bidders *ex ante*, but known *ex post*. To account for differences in bids with common values, the bidders are assumed to possess private information or signals, s_j. These signals are assumed generated from a distribution that is conditioned on the *ex post* common values v. Thus, in a common values setting, the

[24] We use the term purchase auction to refer to auctions where higher bids increase the likelihood of winning. Conversely, procurement auctions are auctions in which lower bids increase the chances of winning.

economic objects of interest are the signals $S = (s_1, \ldots, s_N)$, their joint conditional density $f(S \mid v)$, the common values v, and the marginal density of the common values $f_v(v)$.

To summarize our discussion of auction models: there are three main dimensions along which existing auction models differ:

1. Bidders are uncertain about the *ex post* value of the item.
2. Bidders' private information signals are correlated.
3. Bidders are symmetric in their uncertainties about other bidders' private information.

In an auction where bidders are symmetric, we can summarize the major types of auction models and their primitives in a two-by-two table. Table 1 summarizes the two major differences in the way theorists and empiricists have approached modeling auction bids. The first row of each cell gives the acronym describing the auction; the second and third rows give the information and valuation objects, and the related density functions, that a structural modeler seeks to recover.

Our characterization of the affiliated values (AV) model follows Milgrom and Weber (1982) and McAfee and McMillan (1987). In an AV model, bidders receive private signals $S = (s_1, \ldots, s_N)$ about an item's value and there are also common unknown components v. *Ex ante*, each bidder j is uncertain about the value of the item. Bidders' utilities are (symmetric) functions of the common components v and all bidders' private information signals, S. That is, $v_j = V(s_j, S_j, v)$, where S_j contains all signals but bidder j's. In these models, bidders' private valuations and the common components are assumed affiliated – which loosely means that if a subset of them are large, it is likely the remainder are large.[25] Because the equilibria of affiliated values (AV) models are usually very difficult to characterize, there have been few attempts to estimate general affiliated value models.

Table 1
Private information (conditionally) independent

Uncertainty in final value	YES	NO
YES	PCV $f_v(v)$ $f_{s\mid v}(s_j \mid v)$ $s_1, \ldots, s_N, v$	AV $f(S, v)$ $s_1, \ldots, s_N, v$
NO	IPV $f_s(s_j)$ $s_1, \ldots, s_N$	APV $f_S(S)$ $s_1, \ldots, s_N$

[25] See Milgrom and Weber (1982) for a more complete discussion and references.

IO economists have instead focused on estimating the special cases of the AV model described in the remaining three cells. The bottom row of the table describes two private values (PV) models. In a PV model there is no uncertainty to the bidder's valuation because bidder j observes s_j prior to the auction and thus knows v_j. Bidder j still faces uncertainty in a PV auction, however, because other bidders' valuations are unknown. In an asymmetric independent private values (IPV) model, bidders' presume that the other bidders' values are independently drawn from the marginal densities $f_j(s_j)$. In an affiliated private values (APV) model, nonnegative correlation is allowed. When the bidders share the same beliefs about each others' private valuations, we can represent the density of valuations in a symmetric APV model by $f(s_1, \ldots, s_N)$.

Another special case of interest is a pure common values (PCV) model. In contrast to the private values model, in a PCV model, bidders do not know the value of the item before they bid. All bidders, however, will *ex post* value the item the same. Thus, it is as though there is a single common component v and $V_j(S, v) = V_k(S, v) = v$ for all signals. Such a situation might characterize a situation where bidders are purchasing an item for later resale. To calculate the expected value of winning in a PCV auction, the researcher requires assumptions about the joint distribution of the known signals and the *ex post* value. To facilitate calculations, the usual assumptions are that there is a commonly known prior distribution for v, $f_v(v)$ and that bidders' private information conditional on the signal are (symmetrically) conditionally independent – i.e., $f_{S|v}(S \mid v) = \prod_{j=1}^{N} f_{s|v}(s_j \mid v)$.

We now discuss situations where one can recover the objects listed in this table. The standard approach to developing a structural auction model is to derive equilibrium bid functions for each bidder given each bidder's utility function, the bidder's private signal, other bidders' strategies and the bidder's beliefs about the other bidders' signals. Provided these Bayesian–Nash bid functions are increasing in the unobservable private information and any common values, the empiricist can potentially recover estimates of the unobservables. That is (in a slight abuse of notation), the structural modeler hopes to relate observed data on bids in auction i, B_i, to equilibrium bid function equations: $B_1(s_{1i}), \ldots, B_N(s_{Ni})$. While our notation suggests that the equilibrium bid function for bidder j, $B_j(s_j)$ only depends on the bidder's private information s_j, the equilibrium function also depends on the distribution of private information and common values, $F(S, v)$. This dependence means that in practice we cannot determine the specific form of $B_j(s_j)$ without either (a) making a parametric assumption about $F(S, v)$, or (b) using an estimate $g(B_i \mid Z_i, X_i)$ to recover information on the form of $F(S, v)$.

In nearly all empirical auction models, the process of drawing inferences about the objects in the above table is facilitated by the stark assumption that the only source of error in auction bids are S and v. That is, most empirical models of auctions do not allow for measurement errors in bids or unobserved heterogeneity in the valuation distribution across auctions.[26]

[26] There are exceptions. Paarsch (1992) attempts to model unobserved heterogeneity in timber auctions. Krasnokutskaya (2002) models unobserved heterogeneity in highway procurement auctions.

8.1.3. Nonparametric identification and estimation

Recently, structural modelers have devoted substantial energy to the problem of flexibly estimating the joint density of private information and a single common value component – $f(S, v)$. These efforts reflect the practical reality that the researcher rarely knows *ex ante* what specific $f(S, v)$ bidders used. While this ignorance might seem to favor the researcher estimating general nonparametric density functions for affiliated random variables, such models have proven computationally impractical. This has led researchers to focus on establishing nonparametric identification and estimation results for the special cases described in the previous table.

8.1.3.1. Private values auctions Symmetric independent private values auctions present the simplest identification issues. In a symmetric IPV model, the researcher seeks to recover an estimate of the marginal density of private information $f(s_j)$ (or equivalently, $f(v_j)$) from bid data. The main result in this literature is that data on winning bids are sufficient to nonparametrically identify $f(v_j)$ and estimate v_j. To gain an intuitive understanding of what is involved in deriving nonparametric identification results for private information models, it is useful to begin by considering what happens when there is no private information. By ignoring agent uncertainty and agent beliefs, we can isolate the effect that the econometrician's uncertainty has on inferences. To do this, we compare two procurement auctions.

The following example auction draws an analogy between the problem of how to estimate firm costs in a Bertrand oligopoly setting and the problem of how to estimate the distribution of private values in an IPV setting.

EXAMPLE 10. In a symmetric IPV procurement auction, the bidders' valuations (or in this case costs) are drawn independently from the same marginal distribution $f(c_j)$. Each bidder j only gets to observe their cost c_j. Suppose that each bidder can observe all bidders' costs, $C = (c_1, \ldots, c_N)$ so that each bidder knows the identity of the lowest cost bidder. Because no bidder will find it profitable to bid less than his cost, it is easy to see that in a Nash equilibrium the lowest-cost bidder will win the auction by bidding (slightly less than) the second-lowest cost.

This equilibrium is analogous to what would happen in a homogeneous-product Bertrand oligopoly. In a homogeneous-product Bertrand market where firms have different constant marginal costs, the firm with the lowest marginal cost will end up supplying the entire market at a price equal to the marginal cost of the second-lowest cost firm.

Now consider what an economist could learn by observing equilibrium prices in a set of Bertrand markets. Because the economist knows the equilibrium price equals the marginal cost of the second-lowest cost competitor, they can use a random sample of market prices to estimate the density, $f(c_{[2:N]} \mid X, Z)$, of the second-lowest marginal cost. Moreover, as we shall see shortly, it will be possible under certain assumptions for the economist to recover the density of mar-

ginal costs, $f(c \mid X, Z)$, from $f(c_{[2:N]} \mid X, Z)$. Thus, this example suggests how an economist might recover information about bidders' valuations in an IPV auction *from only data on winning bids*. The key simplifying assumption, which we shortly relax, is that we assumed that the Bertrand competitors were not privately informed about their costs. This makes the solution of the Bertrand and IPV auction very simple in that all but the winning bidder have an incentive to bid their true costs.

Before introducing private information among bidders, it is useful to explore what would happen if the economist had more data than just equilibrium price. For example, suppose that the Bertrand market followed the format of an English button auction. In a descending-price English button auction, all bidders start the auction facing each other with their fingers depressing buttons. The seller then announces continuously lower prices starting from a very high price. Bidders drop out of the auction when they remove their fingers from the buttons. The analogy in a Bertrand market would have prices start out at very high levels with all firms being willing to supply the market. As firms continuously undercut one another, firms would drop out once the price fell to their marginal cost. This process would continue until price hit the marginal cost of the firm with the second-lowest marginal cost. At this point, the firm with the second-lowest marginal cost would drop out and the firm with the lowest cost would supply the entire market at this price. By observing the prices at which all firms dropped out, the economist could directly infer the marginal costs of all but the most efficient firm. Thus, the economist could use the drop-out prices to improve their estimates of the density of marginal costs $f(c_j)$.

This next example considers the effect that correlation among private values has on inferences made from bid data. To do this we compare an APV model to a homogeneous-product, quantity-setting oligopoly model. Again we assume that the oligopoly firms' and bidders' costs are known. Later we will draw a more formal analogy between this oligopoly example and a PV auction.

EXAMPLE 11. Consider an N-firm homogeneous product oligopoly in which firms' constant marginal costs are drawn independently from the joint density $f(c_1, c_2, \ldots, c_N) = f(C)$. Let $P(Q)$ denote the inverse market demand curve, q_i the output of firm i, and let $Q = \sum_{i=1}^{N} q_i$ denote industry output. Assume, as in the previous example, that the suppliers observe C, the vector of marginal costs, before they choose quantity to maximize profits (given the quantity choices of their competitors). The profits of each firm are: $\pi_i(Q) = (P(Q) - c_i)q_i$. The optimal Nash equilibrium quantities solve the N first-order conditions:

$$P = c_i - \frac{\partial P(Q)}{\partial Q} q_i. \tag{97}$$

As we shall see shortly, these first-order conditions closely parallel the first-order conditions that determine equilibrium bids in private value auctions.

Using the implicit function theorem, we can solve these equations for quantities as a function of all firms' costs. Similarly we can use a change-of-variables formula to derive the joint density of $(q_1, q_2, \ldots, q_N)$ from the joint density of $(c_1, c_2, \ldots, c_N)$. Both of these operations require a nonvanishing Jacobian, which amounts to an identification condition for obtaining the joint density of firms costs, $f(c_1, c_2, \ldots, c_N)$, from the joint density of $(q_1, q_2, \ldots, q_N)$. Analogously, in an affiliated private values auction model, there are a set of first-order conditions that relate the privately-observed costs and assumed joint distribution of costs to the optimal bids. [See, for example, Li, Perrigne and Vuong (2002).] By observing a large sample of independent and identical auctions, one could construct an estimate of the joint distribution of the equilibrium bids, $g(B)$. The researcher could then substitute this estimate into the first-order conditions and (provided certain technical conditions are satisfied) use it to recover estimates of the unobserved costs.

Although the above discussion is heuristic, it makes clear that identification hinges on having sufficient similarities in the sampled auctions.[27] As soon as the sampled auctions differ in observable or other unobservable ways, there may be no practical way by which the researcher can reliably recover $f(c_1, c_2, \ldots, c_N)$ from the observed bids.

This point is perhaps easier to appreciate by considering how researchers have estimated the distribution of valuations nonparametrically in an IPV setting. Guerre, Perrigne and Vuong (2000) were the first to devise techniques for recovering a consistent estimate of the underlying distribution of IPV model valuations without making specific parametric assumptions. They model the behavior of N *ex ante* identical risk neutral bidders in a first-price auction. In the case of a procurement auction, this amounts to bidder j maximizing the expected profits

$$E\big[\pi_j(b_1, b_2, \ldots, b_N)\big] = (b_j - c_j)\Pr(b_k > b_j,\ \forall k \neq j \mid c_j), \tag{98}$$

by setting

$$b_j = c_j - \Pr(b_k > b_j,\ \forall k \neq j \mid c_j)\left(\frac{\partial \Pr(b_k \geqslant b_j,\ \forall k \neq j)}{\partial b_j}\right)^{-1}. \tag{99}$$

Here $\Pr(b_k \geqslant b_j,\ \forall k \neq j)$ is the probability that supplier j wins with a low bid of b_j and c_j is bidder j's private cost.

In the symmetric IPV case, the equilibrium bid function simplifies to

$$b_j = \beta(c_j \mid F, N) = c_j + \frac{\int_{c_j}^{\infty}[1 - F(\tau)]^{N-1}\,\mathrm{d}\tau}{[1 - F(c_j)]^{N-1}}, \tag{100}$$

where here we use $\beta(\cdot)$ to denote the equilibrium bid function, $F(c_j)$ is the distribution function of private cost (value) for the item being auctioned. This expression relates the

[27] For illustrative discussions of identification issues see Laffont and Vuong (1996), Guerre, Perrigne and Vuong (2000), and Athey and Haile (2002).

equilibrium bids explicitly to the bidder j's own cost c_j and the distribution function $F(\cdot)$ of all bidders' marginal costs. Because by assumption this expression holds for each of the N bidders across each of the I auctions, the researcher could construct and compare separate estimates of $F(c_j)$ from different random collections of observed bids.

Guerre, Perrigne and Vuong (2000, p. 529) describe a related nonparametric procedure as follows. Assuming no measurement error in the bid data, a straightforward application of the change-of-variables formula yields an expression for the density of each bid b:

$$g(b) = \frac{f(\hat{c})}{|\beta'(\hat{c})|} = \frac{f(\beta^{-1}(b))}{|\beta'(\beta^{-1}(b))|} \tag{101}$$

where $\hat{c} = \beta^{-1}(b)$ is the inverse of the equilibrium bid function, $b = \beta(\hat{c})$, $\beta'(\cdot)$ is the first derivative of $b = \beta(\hat{c})$, and $f(c)$ is the density associated with $F(c)$. Thus, $\hat{c}_j = \beta^{-1}(b_j)$ is the private cost given the observed bid b_j. To apply this formula, they require that the bid function be strictly monotone.

Equation (101) relates the density of observed bids to the unknown density of private costs, apart from the derivative of the equilibrium bid function in the denominator. By differentiating (100) one can obtain an expression for this derivative. Using this expression to substitute out the integral in (100), we obtain

$$\beta(c_j \mid F, n) = c_j + \frac{\beta'(c_j)[1 - F(c_j)]}{(N-1)f(c_j)}. \tag{102}$$

Substituting (101) into this expression and making use of $G(b_j) = F[\beta^{-1}(b_j)]$ gives

$$c_j = b_j - \frac{1}{N-1}\frac{1 - G(b_j)}{g(b_j)}. \tag{103}$$

Here $G(b)$ is the distribution of bids, $g(b)$ is the density of bids, and N is equal to the number of bidders. Thus, to recover the unobserved private costs on the left-hand side, the researcher only requires estimates of the distribution function and density function of bids. Under the assumption that there are no observable or other unobservable differences across auctions, and that $G(\cdot)$ and $g(\cdot)$ are the same across auctions, the researcher can pool data on all bids to estimate $G(\cdot)$ and $g(\cdot)$ nonparametrically. From (103), the researcher can estimate c_j. Once the researcher has estimates of the c_j, nonparametric smoothing techniques can again be used to produce an estimate of the density $f(c)$ or distribution $F(c)$ of private costs.

This same strategy can be used to estimate the density of private values nonparametrically if the researcher only observes winning bids. In this case, Equation (101) must be changed to account for the fact that the winning bid in an IPV procurement auction is that of the lowest-cost bidder. Because the winning bidder has cost $c_{(1:N)}$, the density

of the winning bid b_w is

$$h(b_w) = \frac{\bar{g}(\beta^{-1}(b_w))}{\beta'(\beta^{-1}(b_w))}, \quad \text{where } \bar{g}(z) = N\big[1 - G(z)\big]^{N-1} g(z), \tag{104}$$

and $z = c_{(1:N)}$.

The strength of this nonparametric approach is that it does not require parametric assumptions about unobserved valuations. To see why this flexibility is important economically, it is useful to compare Equation (99)

$$b_j = c_j - \Pr(b_k > b_j,\ \forall k \neq j \mid c_j)\left(\frac{\partial \Pr(b_k \geqslant b_j,\ \forall k \neq j)}{\partial b_j}\right)^{-1} \tag{105}$$

to the standard oligopoly mark-up equation

$$P = c_i - \frac{\partial P(Q)}{\partial Q} q_i.$$

In both equations, the second term on the right-hand side determines the markup over marginal cost. The numerator of Equation (105) is analogous to q_i, the quantity sold. The denominator is the decrease in the probability of winning the auctioned item with an increase in the bid, which is analogous to the decrease in quantity with an increase in price. Just as it was important in oligopoly models to use a demand model that yielded flexible demand elasticities, so too it is important to have a distribution function $F(\cdot)$ that yields flexible bid mark-ups.

There are of course costs to estimating $G(\cdot)$ and $F(\cdot)$ flexibly using nonparametric methods. Chief among them is that the researcher will require data on a large number of similar auctions. In practice the researcher may not be able to reliably estimate $F(\cdot)$ when there are more than a few dimensions of observable auction (X) or bidder (Z) heterogeneities. Moreover, reserve prices introduce the similar truncation issues to those in Hendricks and Porter (1988). Here, truncation of the density typically will require the use of trimming or other data adjustment procedures to obtain an accurate representation of the density close to reserve prices. Subsequent work has explored some of the issues, but substantial problems remain in applying nonparametric techniques to standard auction data sets with differing number of bidders and substantial observed bidder heterogeneity.

The structural modeling literature on auctions also has been concerned with the more general question of whether it is possible to use bid data to discriminate between different private information specifications. Given the analysis above, such questions would seem to be relatively easy to resolve by matching observables to unobservables. For example, it seems at first glance plausible that a general AV model is unidentified because one only has N bids from which to infer the $N + 1$ unobservables – the N costs $c_1, c_2, \ldots, c_N$ and the general valuation v. Laffont and Vuong (1996) were the first to consider this question more formally and establish nonparametric identification results. They showed that for the same number of risk neutral bidders N, that any symmetric

AV model was observationally equivalent to some symmetric APV model. Moreover, they showed that while the symmetric APV and IPV models were nonparamertrically identified, the symmetric common values model was generally unidentified. Athey and Haile (2002) and others have examined the sensitivity of these results to different modeling assumptions and data sets. In particular, several authors have considered whether variation in the number of bidders can add additional identifying information.

8.1.3.2. Pure common value auctions In a pure common value auction, each bidder's private information, s_i, is an imperfect signal about the future value of the item. This additional source of uncertainty introduces another object of estimation – bidders' prior distribution on the value of the item, $f_v(v)$. To see how inferences are made about this prior and the density of private information given v, we consider a common values procurement auction.

In a pure common values procurement auction, all bidders have the same *ex post* cost c of performing a task. By assumption, each bidder j has an unbiased cost signal s_j of the cost of the project. This signal has marginal density $f(s_j \mid c)$ and conditional distribution function $F(s_j \mid c)$. In a PCV model, the bidders' private information signals are assumed conditionally independent and all agents are assumed to have the same prior $f_c(c)$ on the cost c.

In a Bayesian–Nash equilibrium, bidder j chooses b_j to solve the following expected profit maximization problem:

$$\max_{b_j} \Pi(b_j, s_j) = \int_{-\infty}^{\infty} (b_j - c)\big[1 - F\big(\beta^{-1}(b_j) \mid c\big)\big]^{N-1} h(c \mid s_j)\,\mathrm{d}c. \tag{106}$$

In this maximization problem, $b_j - c$ is bidder j's profit from winning, $[1 - F(\beta^{-1}(b_j) \mid c)]^{N-1}$ is bidder j's probability of winning given cost c, and $h(c \mid s_j)$ is the posterior density of c given the signal s_j. Bidder j's posterior density is

$$h(c \mid s_j) = \frac{f(s_j \mid c) f_c(c)}{\int_{-\infty}^{\infty} f(s_j \mid c) f_c(c)\,\mathrm{d}c}. \tag{107}$$

The symmetric Bayesian–Nash equilibrium bid function $\beta_c(s_j)$ is obtained from the first-order condition for the maximization problem. It satisfies the following differential equation

$$\beta_c'(s_j) - \beta_c(s_j) p(s_j) = q(s_j),$$

where

$$p(s_j) = \frac{\int_{-\infty}^{\infty} (N-1)[1 - F(s_j \mid c)]^{N-2} f^2(s_j \mid c) f_c(c)\,\mathrm{d}c}{\int_{-\infty}^{\infty} [1 - F(s_j \mid c)]^{N-1} f(s_j \mid c) f_c(c)\,\mathrm{d}c},$$

$$q(s_j) = -\frac{\int_{-\infty}^{\infty} c(N-1)[1 - F(s_j \mid c)]^{N-2} f^2(s_j \mid c) f_c(c)\,\mathrm{d}c}{\int_{-\infty}^{\infty} [1 - F(s_j \mid c)]^{N-1} f(s_j \mid c) f_c(c)\,\mathrm{d}c}.$$

This is an ordinary linear differential equation with solution

$$\beta_c(s_j) = \frac{1}{r(s_j)}\left[\int_{-\infty}^{s_j} r(u)q(u)\,\mathrm{d}u + k\right], \quad \text{with } r(\tau) = \exp\left(\int_{-\infty}^{\tau} p(u)\,\mathrm{d}u\right). \tag{108}$$

The constant k is determined by boundary conditions.

At first it might seem, following our discussion of the IPV model, that a researcher could use these integral equations as a basis for nonparametric estimation. Closer inspection of the differential equation reveals that for a given bid function, $\beta_C(s_j)$, there are a number of distribution functions $f(s_j \mid c)$ and $f_c(c)$ that could satisfy the above differential equation. This is in fact the nonidentification result of Laffont and Vuong (1996).

8.2. *Further issues*

These discussions of nonparametric identification show that identification can hinge delicately on any of several stochastic and economic assumptions. Indeed, there remain a great many combinations of auction formats and assumptions yet to be explored in the literature. For example, there are few general results on what can be identified with risk aversion. What results we do currently have suggest that much stronger identifying assumptions will be required when bidders are risk averse. [See Campo et al. (2003).]

It also is important to realize that most auction models in the theoretical and empirical literatures maintain that bidders' beliefs are symmetric. When bidders' beliefs differ for observable and unobservable reasons, auction models become much more challenging – both because it is more difficult to compute pure-strategy equilibrium bids and because there may be no pure-strategy equilibrium bids.

There also remain many institutional details that have yet to be fully explored in the nonparametric identification literature. For example, the presence of reserve prices can complicate both equilibrium bid functions and nonparametric estimation. These complications can destroy the identification of part or all of the relevant distributions of signals and common values. Another important assumption that may not hold in practice is the assumption that the number of bidders N is exogenous and known by the researcher. In many auctions, there appear to be few limitations on who can bid. One reason presumably why we do not see hundreds of bidders is because many are confident that their probability of winning is sufficiently low that this does not justify the expense of preparing and submitting a bid. Additionally, potential bidders could be deterred by the knowledge that other bidders are participating in the auction.

Despite all these limitations, nonparametric identification results and nonparametric estimation methods provide a useful reference for understanding what can be identified by imposing minimal economic rationality on observed bids. We now briefly consider what additional information can be gained by imposing parametric structure.

8.3. Parametric specifications for auction market equilibria

The previous subsection showed how structural econometric modelers have used first-order conditions from static auction models to estimate the primitives of alternative auction models. These first-order conditions had the form

$$b_j = \beta_j(v_j, N, F_j), \tag{109}$$

where v_j was a private valuation or signal, N is the number of potential or actual bidders, and $F_j(\cdot)$ was a joint distribution of private information and common uncertainties. From this equation we see that it is both bidder j's private valuation or signal v_j as well as bidder j's beliefs about other bidders' private information and common uncertainties, $F_j(\cdot)$, that affect observed bids. Under certain (identification) conditions, the system of bid functions can be inverted to recover the v_j

$$\begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix} = \begin{bmatrix} \beta_1^{-1}(B, N, F_1, \ldots, F_N) \\ \vdots \\ \beta_N^{-1}(B, N, F_1, \ldots, F_N) \end{bmatrix}$$

but only provided the bidder beliefs, F_j, are known or can be consistently estimated.

Because in a typical application F_j is unknown, it seems highly desirable that empiricists be as flexible as possible when estimating F_j. As we indicated repeatedly above, this desire raises a paradox: the cost of statistical flexibility may be economic flexibility. For example, to even begin to apply nonparametric techniques we must impose symmetry, $F_j = F_k$. Further, researchers typically do not have sufficient data to estimate general F's when N varies considerably across auctions or when there are many variables that enter the bid function (109). For this reason alone, many researchers have been willing to entertain parametric specifications for F. There are additional reasons to favor parametric specifications. One important one is that parametric specifications can identify economic quantities that are nonparametrically underidentified.

Some empirical researchers feel that as a matter of principle if something is not identified nonparametrically, one should never make parametric assumptions to identify it. Other researchers favor such restrictions if they lead to useful parameter estimates or counterfactuals. We hope it is clear by now that our position is that it is acceptable to make parametric assumptions as long as these assumptions are economically sensible and do not contradict the data. To appreciate the trade-offs that can arise in adding parametric structure, it is useful to see the trade-offs that Paarsch (1997) considered when developing a structural model of British Columbia government timber auctions. Paarsch's goal was to estimate a model of open-outcry timber auction within which he could ask whether the observed government reserve prices were revenue-maximizing. This is an ideal setting in which to use a structural model, because Paarsch seeks to perform counterfactual comparisons.

Several practical realities prevent Paarsch from employing the nonparametric estimation procedures discussed in the previous subsection. First, Paarsch has data on fewer

than 200 auctions. With less than 200 auctions, he has little hope of obtaining sensible estimates of a high dimensional conditional bid density. Second, there are at least five important observable dimensions along which the timber tracts differ. These differences include: the species composition of the tract, the amount of each species on the tract, the distance of the tract to local mills, and potential nonlinearities in harvesting costs. Third, the presence of a reserve price in the timber auctions introduces the same truncation problems present in Hendricks and Porter's (1988) descriptive study of offshore oil and gas lease auctions. Because Paarsch does not observe bids below observed reserve prices, he cannot estimate $F(\cdot)$ nonparametrically in these regions and thus cannot evaluate the revenue consequences of lowering reserve prices. Fourth, although his original sample consists of open-outcry and sealed-bid auctions, he chooses to focus exclusively on the open-outcry auctions. In open-outcry auctions, the dynamics of bidding can affect the observed sequence of bids.

Individually and collectively, these practical realities force Paarsch into a parametric specification of bidders' valuations and harvesting costs. Moreover, these realities also appear to force stark assumptions in order to obtain an estimable model. For instance, while Hendricks and Porter's discussion of oil and gas leases might suggest timber auctions have common value components, Paarsch rules out this possibility for at least two reasons. First, as a practical matter Paarsch's previous work showed that the private value auction framework provided as good or better explanation of winning bids as a pure common value auction framework. Second, English auctions in which bidders have common values are much more difficult to model. In an open-outcry English auction in which bidders' values are affiliated, bidders will revise their beliefs and bids according to the bidding history – not just their own private signal. For these and perhaps other reasons, Paarsch is led to adopt an independent private values framework.

In addition to adopting a private values framework, Paarsch also adopts a particular auction format to model the open-outcry structure of his timber auctions. Specifically, he assumes bid data are recorded via a "button" English auction. In an English button auction, all bidders begin the auction in plain view of one another with their fingers on a button. They listen to the auctioneer continuously call out increasing prices starting from the reserve price. A bidder exits the bidding (permanently) by removing their finger from the button. The last bidder depressing the button wins the auction at a price equal to the second-to-last bidder's value (or cost).

It is not hard to see why Paarsch makes this assumption. Because bidders' valuations are independent and private, bidders do not update their beliefs about valuations during the bidding. Moreover, their equilibrium strategy is to stay in the auction until the bidding reaches their valuation, at which point they drop out. (The winning bidder of course drops out when the second-to-last bidder does.) How does this equilibrium map into Equation (109)? Now, for all losing bidders

$$b_j = \beta(v_j, N, F_j) = v_j. \tag{110}$$

Thus, Paarsch's IPV assumption, when combined with the button English auction assumption, allows Paarsch to recover the valuations of all but the winning bidder from

the observed bids. There is little to do in terms of estimation. From any and all of the losing bids it would be possible to estimate a symmetric distribution of private values $F(\cdot)$ nonparametrically were it not for the fact that $F(\cdot)$ is conditioned on a large number of observables that vary across auctions.[28]

The automatic recovery of values from bids should sound familiar. This was exactly the solution in Example 10 where we drew a parallel between a perfect-information Bertrand model and an IPV model in which all bidders knew all costs. Here, Paarsch can recover exactly the valuation (here, profits) of the second-highest bidder. Paarsch also observes the third-highest valuation, and so on. Thus, if Paarsch were only interested in recovering valuations from bids, he could effectively dispense with the private information assumption altogether. To perform his reserve price counterfactuals, he could simply treat $F(\cdot)$ as a statistical construct that captures not private information of the bidders but simply unobserved (to him) reasons why bidders' profits would differ across tracts.

Other aspects of Paarsch's application have a bearing on how Paarsch estimates and interprets $F(\cdot)$ however. One of these is the match between the button auction model and the way the auctions were run and data collected. In a button auction, the last "bid" of a bidder is the price at which the bidder removes their finger from the button. In an open-outcry auction, the last observed bid is the last oral bid. For a variety of reasons, bidders may space their bids in an open out-cry auction, yielding the possibility of nonuniform jumps in bids. If this is the case, it is unclear how one should interpret the last bids in Paarsch's data.

There are other features of the timber auctions that affect the empirical model. To appreciate these features, it is useful to go into the details of his application. Paarsch decomposes each bidder j's valuation v_{ij} into an average revenue per tree on the tract, r_i, and average harvesting costs, c_{ij}. That is, $v_{ij} = r_i - c_{ij}$. The absence of a bidder j subscript on revenues, and the lack of any common value component in the auction, immediately implies that revenues are known to all bidders. In addition, Paarsch assumes that he observes the revenues bidders observe. He calculates these revenues as a sum of species prices times the amount of each species the government estimates is on each tract. Thus, when it comes to distinguishing between private information or unobserved heterogeneity models, it is the individual differences in harvesting costs that are important.

A key novelty of Paarsch's paper is that he models the difference between the potential number of bidders in an auction and the number who end up bidding. To see why this distinction is important, notice that the reserve prices in Paarsch's auctions truncate not only the distribution of observed bids, but lead to a difference between the potential number, $\mathbf{N_i}$, and actual number, N_i, of bidders. To model this difference, Paarsch makes parametric assumptions about the distribution of bidders' harvesting costs and how they

[28] There is additional information in the condition that the winning bidder's valuation exceeds the winning bid. Paarsch could presumably use this inequality to improve the precision of estimates of $F(\cdot)$.

vary across auctions. Specifically, he introduces two types of bidder heterogeneity. In the leading case, he models a bidder's average cost c_{ij} as being drawn from a Weibull density

$$c_{ij} \sim F(c \mid \delta) = 1 - \exp\bigl(-\delta_1 (c - c_{\min})^{\delta_2}\bigr), \quad c \in [c_{\min}, \infty],$$

where the δ's are unknown parameters that affect the heterogeneity of costs and $c_{\min}$ is a lower bound equal to

$$c_{\min} = \gamma_0 \frac{1}{q} + \gamma_2 q + \gamma_3 q^2 + \gamma_4 d.$$

Here, q measures the size of the tract and d is the distance to the closest timber mill. The δ and γ parameters help capture reasons why the distribution of average harvesting costs might vary across auctions. Adequately capturing such variation would be crucial to accurate counterfactual calculations. In a second specification, Paarsch considers the consequences of assuming fixed harvesting costs, γ_0, are random.

By modeling costs parametrically, Paarsch can use maximum likelihood to estimate the unknown costs of bidders. These costs are critical to his optimal reserve price calculations. One major problem remains however – how to account for the fact that while he observes the number of bidders, N_i, he does not observe the number of potential bidders, $\mathbf{N_i}$.

To appreciate this problem, consider timber auctions that have at least 2 bids, i.e., $N_i \geqslant 2$. There will be a difference between N_i and $\mathbf{N_i}$ when potential bidders with extremely high harvesting costs find it unprofitable to bid above the reserve price. The likelihood of observing the order statistic data $c_{[2:\mathbf{N_i}]}, c_{[3:\mathbf{N_i}]}, \ldots, c_{[N_i:\mathbf{N_i}]}$ and N_i is

$$L(\gamma, \theta \mid C_i, N_i) = \binom{\mathbf{N_i}}{N_i} \bigl[1 - F(c^*)\bigr]^{\mathbf{N_i} - N_i} F(c^*)^{N_i} \times N_i! \frac{F(c_{[2:\mathbf{N_i}]})}{F(c^*)^{N_i}} \prod_{j=2}^{N_i} f(c_{[j:\mathbf{N_i}]}). \tag{111}$$

The first portion of this likelihood function (before the $\times$) is the (binomial) probability of observing $\mathbf{N_1} - N_i$ cost draws below the cost c^*, where c^* is the cost that would result in profit of zero at the reserve price. The second portion is the density of observable average harvesting costs given N_i and that the unobserved lowest cost satisfies $c_{[1:N]} < c_{[2:N]}$. The problem with trying to estimate δ and γ with this likelihood function is that Paarsch cannot compute the likelihood function unless he knows the number of potential bidders $\mathbf{N_i}$ for each auction. Because he does not know $\mathbf{N_i}$, he could treat each $\mathbf{N_i}$ as a parameter to be estimated. This, however, amounts to introducing a new parameter for each auction. As is recognized in the econometrics literature, the introduction of so many parameters will make the maximum likelihood estimator inconsistent. Absent a solution then to this problem, Paarsch, and many later auction researchers, are stuck.

A main contribution of Paarsch's paper is to show that a conditional likelihood function approach [Andersen (1970)] can be used to obtain consistent estimates of δ and γ. The conditional likelihood approach works as follows. Let $f(C_i, N_i \mid \mathbf{N_i})$ be the joint density of observed costs and bidders conditional on the unobserved potential number of bidders in auction i. According to the conditional maximum likelihood approach, if this density can be factored into two pieces of the form

$$f(C_i, N_i \mid \mathbf{N_i}, \delta, \gamma) = g(N_i \mid \mathbf{N_i}, \delta, \gamma) \times h(C_i \mid N_i, \delta, \gamma),$$

then one can obtain consistent estimates of δ and γ by maximizing the conditional likelihood function $h(C_i \mid N_i, \delta, \gamma)$. Paarsch's contribution is to show that for this specific IPV auction, the likelihood function (111) has this form, with N_i serving as a sufficient statistic for the unknown potential number of entrants.

We now are in a position to return to the point on which we began this example. While the costs of parametric assumptions in many applications are self-evident, the benefits are sometimes less clear. One important benefit of parametric structure is that it may allow the researcher to identify a quantity of interest. In Paarsch's case, the realities of timber auctions necessitated several strong modeling and parametric assumptions, such as private values and an English button format. On the other hand, the resulting model did overcome a significant handicap, which is that the number of potential bidders is rarely known.

Whether this benefit justifies the starkness of the assumptions, has to be viewed from at least three vantages. First, is the benefit practically useful? The answer here appears to be a resounding yes. Without it Paarsch could not estimate his model and perform the counterfactual optimal reserve price calculations. Second, does the parametric structure deliver the end result? In Paarsch's case, the answer is unclear. Finally, does the additional structure adequately capture the economics of the agents' behavior, particularly when it comes to the counterfactuals? To answer this question, Paarsch tries to convince readers by reporting alternative models and estimates.

8.4. Why estimate a structural auction model?

Previously, we asserted that researchers should not attempt a structural model without a convincing explanation of how its benefits will outweigh potentially restrictive and untestable assumptions. This advice seems particularly relevant when considering how to model auction bid data.

The main benefit of a structural auction model would seem to be that it allows the researcher to estimate the distribution of bidders' valuations (or similar objects). Such estimates can in principle be used to evaluate an auction's efficiency or how changes in the rules would affect the seller's revenues.

In actual applications, however, these benefits are only achieved at the cost of restrictions on bidders' information. In particular, the vast majority of structural auction models either exclusively estimate independent private values or pure common values models. The reasons for this specialization are not too hard to find – more realistic

affiliated models are analytical and computationally intractable.[29] Restrictions on the distribution of bidders' information naturally limit the applicability of the estimated model. For example, it makes little sense to estimate an IPV model and then use those estimates to model what would happen if there was a common value.

Even if we are willing to accept the independent private values or pure common values assumptions, there are other factors that can affect the value of structural estimates. Consider what we learn from estimating an IPV model. The best one can hope for is to be able to recover a precise nonparametric estimate of the distribution of bidder valuations $F(v_j)$ above for valuations that would lead to bids above the reserve price. But what is the value of knowing $F(v_j)$? We believe that the answer is that there is little or no value unless we can somehow say that $F(v_j)$ is applicable to past auctions or future auctions. For example, we could imagine estimating $F(v_j)$ during a period in which there was no collusion among bidders and then trying to use the estimated density to compare bids (valuations) when bidders were perhaps colluding. Alternatively, like Paarsch (1997), one could perform counterfactuals that involve changing some aspect of the auction like the reserve price.

The key question is: How does one know that the estimated valuation distribution is relevant to other auctions? Our position is that to be convincing, the structural modeler has to have a convincing explanation for when $F(v_j)$ is likely or unlikely to change from auction to auction. To take Paarsch's (1997) timber auction model as an example, we might ask: When would his estimates be relevant for a timber auction in another Canadian province? To answer this question, we ultimately would need to understand how timber auctions are different. This is not a question that auction theory itself can answer directly. Instead, the answer likely lies in the specifics of what is being auctioned and how it is auctioned. Thus, we see that economic theory often can only go so far in answering specification issues. In the end, the econometrician will have to pick and justify conditioning variables. Ideally, these choices will be made with the aid of economics, but in practice it is knowledge of the industry, institutions and data that will likely make the analysis convincing.

Suppose we can accept the assumptions of a structural auction model, what can we do with the resulting estimates? Structural auction models can in principle facilitate useful counterfactual experiments. Paarsch's (1997) evaluation of optimal reserve prices is one example. Other researchers have used structural models to evaluate alternative winning bid rules. One area where structural auction models have yet to make much headway is in diagnosing bidder collusion. Here there are two problems. First, economists do not have particularly good models of how colluding bidders behave. Indeed, the modeler often is confronted with the paradox: rationality suggests that colluding bidders will scramble their bids so as to make detection extremely difficult. To date, most of what

[29] Although there have been some attempts to compare private values and common values models, these tests invariably rest heavily on functional form and other assumptions. In the end, little progress has been made using structural models to decide the appropriateness of different information structures.

structural models have contributed to the detecting collusion literature are benchmark noncooperative models of bids. IO economists have used these models to look for suspect bid clustering, skewing or correlation.

There are additional practical issues that limit the usefulness of structural auction models. One set pertains to dynamic considerations. In many applications, the data come from repeated auctions where the same bidders bid against one another. This repetition raises two issues. First, in repeated auctions, bidders' subsequent valuations may be influenced by the number of units they have won in the past. In this case, symmetric information models no longer make sense. Second, in repeated auctions bidders likely will internalize information and strategic externalities that their bids today may have for bids tomorrow.

8.5. *Extensions of basic auctions models*

Recent research has addressed many of the limitations associated with the auction frameworks described in Table 1. It is well beyond the scope of this chapter to even begin to survey this literature. Interested readers should consult Hendricks and Porter (in press).

There are some developments that fit in with our earlier discussions that are worth noting briefly. Laffont, Ossard and Vuong (1995), for example, extended the IPV paradigm to allow for both observable and unobservable heterogeneity across auctions. Although their estimation procedure assumes a parametric model for the distribution of private valuations, they devise a clever estimation technique based on simulated nonlinear least-squares that does not require them to compute the equilibrium bid functions. Instead their technique simulates the expected value of the winning bid for an arbitrary distribution of private values and a potentially binding reserve price. They also treat the number of potential bidders as a random variable.

Haile and Tamer (2003) explore the empirical implications of English auctions. The button English auctions we considered earlier are a special type of English auction. In open-outcry English auctions, bidders can bid whenever they are willing to best the outstanding bid (plus any minimum bid increment). Exactly what order and when bidders will bid is something left to the auction's format and strategic considerations. In general, the dynamics of English auctions are extremely difficult to analyze. Rather than try and detail the dynamics of the bidding, Haile and Tamer take a minimalist approach by using potentially weak restrictions on players' bids. Specifically, Haile and Tamer maintain that observed bids need only satisfy the following two restrictions: (1) bidders do not bid more than they are willing to pay; and, (2) bidders do not allow an opponent to win at a price they are willing to beat. Using these assumptions, they derive bounds on the distribution of valuations and bids above reserve prices. These bounds become exact for a button auction and are weak bounds for other English auctions.

There has also been recent empirical research on multi-unit auctions. Virtually all wholesale electricity markets operating around the world run daily multi-unit auctions to determine which generation facilities are able to supply energy. Each day suppliers

submit nondecreasing step functions expressing their willingness each hour to supply electricity for the next 24 hours. The system operator then computes the least cost way to meet demand in each hour based on these bids. Wolak (2000) develops a model of expected profit-maximizing bidding behavior in such markets. Wolak (2003) uses this model to estimate bidder cost functions. He shows that because of the richness of the bid functions that market participants submit, the only assumption required to recover these cost function estimates is expected profit maximizing bidding behavior. An important difference between these multiple-good auctions and single good auctions is that in a multi-unit auction, suppliers compete over how many units they sell. Consequently, residual demand (market demand less the willingness to supply functions of all other market participants) is observable *ex post*, and this provides the information necessary to identify the supplier's underlying marginal cost function.

As should be clear from this brief discussion, significant progress has been made in deriving flexible modeling frameworks which allow empirical IO researchers to recover information about the distribution of private information in auction models under minimal assumptions.

9. Games with incomplete information: Principal-agent contracting models

Recently, IO economists have begun to develop structural econometric models of regulator and regulated firm interactions. These empirical models are more ambitious than the auction or oligopoly models discussed in the previous sections. Similar to oligopoly models but unlike auction models, these models seek to estimate production and demand functions. Similar to auction models but unlike most oligopoly models, these models seek to account for the impact of asymmetric information on agents' strategic interactions. These ambitious modeling goals usually require the researcher to rely on stronger parametric and distributional assumptions to identify and estimate economic primitives. The main goal of this section is to discuss why models of regulatory interactions require this structure.

As in auctions, private information plays a critical role in regulatory proceedings. IO economists have recently used principal-agent contracting models to characterize regulatory proceedings in which regulators set the prices (or "rates") regulated firms (or "utilities") charge. A key insight from these models is that when a utility has superior information about the underlying economic environment, it can exploit that information to earn greater profits than it would if the regulator were equally informed. This paradigm for studying regulator–utility interactions has received such widespread acceptance among IO economists that Laffont and Tirole (1993) have coined the phrase "new regulatory economics".

One of the most important economic primitives in these contracting models is the economist's specification of the regulated firm's private information. The two main types of private information a utility can have are private information about its production process or its demand. The regulated firm has no incentive to reveal this private

information to the regulator because the regulator would use this information against them in the rate-setting process. The regulator in turn is aware that the firm has private information, and takes this into account in setting rates. Economic theorists model this interaction by computing optimal "second-best" solutions to a revelation game. In this game, the regulator announces a price schedule and a transfer payment that are functions of the firm's reported private information.

A critical constraint on the firm is that it must serve all demand consistent with the private information it reports. (Its private information determines the price granted by the regulator.) Under an optimal "second-best" solution, the price schedule chosen by the regulator maximizes a social welfare function subject to the constraints that: (1) the firm finds it profit maximizing to report its true private information to the regulator; and, (2) the firm expects to earn profits sufficient to keep it from exiting the industry. Although these theoretical models are stylized static depictions of regulatory interactions, they do capture important features of the asymmetric information problem faced by actual regulators. Important examples of this work include Baron and Myerson (1982), Baron and Besanko (1984, 1987), Besanko (1984) and Laffont and Tirole (1986).

Historically, empirical IO economists have largely ignored the impact of regulated firms' private information on both regulated firm and regulator behavior. Instead, empirical IO economists have estimated conventional cost and demand functions using standard cost functions, factor demand equations and product demand models. Christensen and Greene's (1976) study of electric utility costs is a classic example of regulated firm cost function estimation. Evans and Heckman (1984) provide a more recent example of cost function estimation applied to the AT&T divestiture decision. In virtually all cost function studies, statistical tests of cost-minimizing behavior are rejected.

The rejection of cost-minimizing behavior is not too surprising if one recognizes the presence of private information. A regulated firm with private information need not find it profit-maximizing to minimize costs if it can distort its behavior to obtain better prices from the regulator. Given these incentives, estimation procedures that assume cost minimization behavior will yield inconsistent estimates of the underlying economic primitives.

The remainder of this section follows the format of the previous section. First, we describe the data a researcher has in a typical application. We then develop a simple model that illustrates what economic primitives can be recovered from these data. After considering nonparametric identification, we discuss the practical limitations of nonparametric identification results. This then leads us to describe how parametric assumptions can be used to identify economic primitives. We illustrate this discussion using Wolak's 1994 study of Class A California Water Utilities. We close with a short discussion of subsequent related empirical work.

9.1. Observables and unobservables

Empirical research on regulated industries benefits from regulatory proceedings that make rich cost and revenue data publically available. On the cost side, for example,

regulated utilities typically must report detailed data on inputs, X, and input prices, p_X. Inputs consist of information on the firm's capital (K), labor (L), energy (E) and materials (M) choices associated with an observed output Q. Additionally, the researcher also will have information on input prices, $p_X = (p_K, p_L, p_E, p_M)'$. Using these data, a researcher can construct an estimate of the total cost, C, of producing the observed output level Q. In terms of the above notation

$$C = p_K K + p_L L + p_E E + p_M M. \tag{112}$$

On the output (or revenue) side, firms provide both retrospective and prospective quantity and revenue data. The prospective quantity data reflect the reality that the regulator must set prices before either it or the firm know what demand will be. When setting price, the regulator attempts to balance two competing goals: (1) it must allow the firm to recover all "prudently" incurred costs; and, (2) it must provide strong incentives for the firm to produce in an efficient manner. To model the prospective nature of the regulator's pricing decisions, it is imagined that demand equals $D(p_Q, Z, \epsilon_Q) = Q$, where p_Q is an output price set by the regulator, Z is a vector of observable variables assumed to shift demand and ϵ_Q is an unobserved demand shifter.

Regulatory models differ according to whether ϵ_Q is known to the firm (i.e., is private information) or is unknown to the firm before the firm reports to the regulator. In what follows, we only explore models in which the firm has private information about its production function. Thus, ϵ_Q here does not reflect private information. The econometrician of course never observes ϵ_Q.

Given these cost and output data, all an empirical researcher can do is consistently estimate the joint density of regulated prices, firm outputs, firm input choices, and total costs – conditional on input prices (p_X) and any demand shifters (Z); i.e., the researcher can estimate $h(p_Q, Q, X, C \mid p_X, Z)$. Input prices and the demand observables are used as conditioning variables because firms are thought to be unable to impact input prices or factors that influence demand. Thus, these vectors Z and p_X are usually assumed to be distributed independently of all of the unobservables in the econometric model.

To obtain a consistent estimate of the firm's production process, the researcher must be very specific about how the utility's private information interacts with the regulatory process. In what follows, we explore models in which the regulated firm has private information about its production process. We restrict our attention to private information on the production side in keeping with Wolak's (1994) empirical model. Specifically, we model the firm's private information as a single parameter that enters the firm's production function $Q = f(K, L, E, M, \theta)$. The firm knows θ from the start and all the regulator knows at the start is the density of θ, $f_\theta(\theta)$, where $\theta \in [\theta_l, \theta_h]$. Absent further assumptions on the distributions of θ and ϵ_Q, and specific functional forms for $D(p_Q, Z, \epsilon_Q)$ and $f(K, L, E, M, \theta)$, little or nothing can be deduced about these underlying economic primitives from $h(p_Q, Q, X, C \mid p_X, Z)$. This is because the firm's input choices will depend in an unknown way on θ, which implies that total cost, C,

does as well. Additionally, because the firm must by law satisfy all demand at the regulated price, the firm's output will depend on the realization of ϵ_Q, the unobservable demand shifter. This implies that the firm's input choices and total cost will also be a functions of the realization of ϵ_Q. Consequently, without functional form restrictions on the demand and production functions, or assumptions about the forms of the distributions of θ and ϵ_Q, the researcher will be unable to identify demand and cost functions from $h(p_Q, Q, X, C \mid p_X, Z)$.

These observations lead us to consider the types of functional form and distributional assumptions that can lead to identification. We will see that nonparametric identification of the distribution of private information, as in independent private values auction models, hinges on a monotonicity condition. We show that strong economic or statistical assumptions are required to guarantee monotonicity. We then discuss parametric models. These models rely on functional form and distributional assumptions to identify the underlying economic and information primitives.

9.2. *Economic models of regulator–utility interactions*

Baron (1989) provides a useful theoretical model for thinking about empirical models of regulator and utility interactions. He assumes $C(q, \theta) = \theta q + K$ where θ is the firm's private marginal cost of producing output q. No explicit economic rationale is provided for the cost function. In particular, there is no reason to believe that the firm produces its output at minimum cost for any value of θ. In this sense, we can think of $C(q, \theta)$ as a behavioral cost function; it gives the cost of producing output q given θ.[30] Additionally, Baron assumes $D(p)$ represents the quantity demanded at the regulated price p. Thus, in his model there is no demand uncertainty.

In Baron's model, the regulator fixes a price schedule, $p(\theta)$, and a monthly (or annual) fixed fee schedule, $T(\theta)$, that give prices and fixed fees as a function of the firm's announced marginal cost θ. Given the price and fixed fee schedules, the firm announces a marginal cost, $\hat{\theta}$, to maximize its profits

$$\pi(\hat{\theta}; \theta) = p(\hat{\theta}) D\big(p(\hat{\theta})\big) + T(\hat{\theta}) - \theta D\big(p(\hat{\theta})\big) - K. \tag{113}$$

There are two constraints imposed on the regulator's price and fee optimization problem. The first is a truth-telling or incentive compatibility constraint. This constraint requires that a firm of type θ will report its true type. In other words, a truthful report must yield the firm profits that are greater than or equal to profits it could obtain through any other feasible report in the support of θ. Mathematically, this implies:

$$\pi(\theta) \equiv \pi(\theta; \theta) \geqslant \pi(\hat{\theta}, \theta), \quad \forall \hat{\theta} \in [\theta_l, \theta_h] \text{ and } \forall \theta \in [\theta_l, \theta_h]. \tag{114}$$

[30] By behavioral cost function we mean only that the firm behaves according to a consistent set of rules that yield this stable relationship between costs and q for a given value of θ. One possible set of behavioral rules is to minimize total production costs, but this is not necessary because, as discussed above, the firm may have little incentive to produce its output according to minimum cost.

As Baron notes, these constraints are global. That is, they must be satisfied for each θ and all feasible reports $\hat{\theta}$.

The second constraint is called the participation constraint or individual rationality constraint. It states that regardless of the firm's true value of θ, it must receive more than its outside option. Here this means that the firm must earn nonnegative profits. Mathematically,

$$\pi(\theta) \geqslant 0, \quad \forall \theta \in [\theta_l, \theta_h]. \tag{115}$$

Because it is extremely complicated to impose the global truth-telling constraint on the regulator's optimization problem, theorists typically make assumptions about economic primitives so that satisfaction of local truth-telling implies satisfaction of global truth-telling. These assumptions are analogous to those in auction models that make the bid functions monotone in the bidders' valuations.

Baron (1989, pp. 1366–1367) shows that the local truth-telling constraint for this problem is the following differential equation in θ:

$$\frac{\mathrm{d}\pi(\theta)}{\mathrm{d}\theta} = -C_\theta\big(D\big(p(\theta)\big), \theta\big). \tag{116}$$

This equation tells how profits must increase as a function of θ in order to induce local truth telling. In words, for small deviations from truthful reporting, the firm experiences a decline in profits. This condition can easily be checked for the assumed cost function, as $C_\theta = q > 0$ for all θ.

As Baron notes, Equation (116) can be integrated to produce an expression for the firm's profit

$$\pi(\theta) = \int_\theta^{\theta_h} C_\theta\big(D\big(p(x)\big), x\big)\,\mathrm{d}x + \pi(\theta_h). \tag{117}$$

This equation implies that the participant constraint can be simplified to

$$\pi(\theta_h) \geqslant 0, \tag{118}$$

which means that the least efficient firm, as parameterized by θ, must earn nonnegative profits. Using the definition of $\pi(\theta)$ in Equation (114), we can re-write Equation (113) as

$$\pi(\theta) = p(\theta)D\big(p(\theta)\big) + T(\theta) - \theta D\big(p(\theta)\big) - K. \tag{119}$$

Deriving the optimal price and fixed fee functions requires specifying the regulator's objective function. The general objective function considered for the regulator is a weighted sum of consumer and producer surplus. Because both consumer and producer surplus will depend on the firm's actions, which depend on the unobserved θ, the regulator must use its knowledge of $f(\theta)$ to compute an expected surplus function

$$W = \int_\theta^{\theta_h} \left[\int_{p(\theta)}^{\infty} D(x)\,\mathrm{d}x - T(\theta) + \alpha\pi(\theta) \right] f(\theta)\,\mathrm{d}\theta, \tag{120}$$

where α is the relative weight given to the firm's profits in the regulator's objective function. The regulator is assumed to choose the price and fixed fee schedules to maximize (120) subject to (115), (116), and (119) using calculus of variations techniques.

Baron (1989) shows that the optimal price schedule takes the form

$$p(\theta) = \theta + (1-\alpha)\frac{F(\theta)}{f(\theta)}, \tag{121}$$

which looks very similar to Equation (99) in the independent private values auction case, apart from the parameter α. Baron shows that a sufficient condition for satisfaction of the local truth-telling constraint to imply satisfaction of the global truth-telling constraint is that $p(\theta)$ is nondecreasing in θ. This equation shows that monotonicity of the price function imposes restrictions on the distribution of θ. Specifically, if $F(\theta)/f(\theta)$ is nondecreasing in θ, then $p(\theta)$ is nondecreasing in θ.

If the value of α is known to the econometrician and firms face the same cost function and nonstochastic demand function, then it is possible to recover a consistent estimate of the density of θ, $f(\theta)$, from prices. Such an exercise would follow the "change-of-variables" logic applied to the first-order condition in sealed-bid IPV auction models. It is important to emphasize all of the assumptions necessary for this identification result. Besides assuming firms have the same cost function and density of private information, we have assumed the demand function is the same across all observations. In other words, $D(p)$ cannot vary across observations and there are no unobservable ϵ_Q or observable demand shifters. Equation (121) also depends crucially on the functional form of $C(q,\theta)$. Without the constant marginal cost assumption, the regulator's optimal price schedule will depend on the demand function $D(p)$.[31]

Although this nonparametric identification result may at first seem appealing, it should not give much comfort to regulatory economics researchers for three reasons. First, it is difficult to imagine circumstances where the researcher will know the value of α. Second, the underlying cost function tells the researcher nothing about the technology of production. As noted earlier, $C(q,\theta)$ simply characterizes the relationship between production costs, q, and θ. The researcher cannot say anything about the returns to scale in production, the elasticity of substitution between inputs or the extent to which the regulatory process results in deviations from minimum cost production. Moreover, the manner in which θ enters the cost function is extremely restrictive. Third, nonparametric identification rests on unrealistic assumptions about the extent of observed and unobserved heterogeneity in the production process and demand. Specifically, in this model the only reason market prices differ across observations is because of different realizations of θ. It is difficult to imagine a sample of regulator–utility interactions with no observed or unobserved heterogeneity in the production and demand functions.

[31] See the discussion of Wolak (1994) below. In addition to recovering the density $f(\cdot)$ nonparametrically, it is possible to recover a consistent estimate of K from information on the regulated quantity and total production cost. Also, if the researcher is willing to assume $D(p)$ is the same for all observations in the sample, then the set of observed (p_Q, Q) pairs will nonparametrically trace out the demand curve $D(p)$.

Some of these shortcomings can be overcome by explicitly specifying an underlying production function and how it depends on θ. The firm's observed cost function can then be derived from the assumption of expected profit-maximizing behavior subject to the constraints imposed on firm behavior by the regulatory process. Both observed and unobserved heterogeneity can also be allowed in the production function and the demand function facing the regulated firm. However, there is a cost of being more general – nonparametric identification is lost, just as it is in the case of auction models. As we now show, however, by clever choices of functional forms and distributional assumptions, the researcher can estimate a rich set of underlying economic primitives.

9.3. Estimating productions functions accounting for private information

Wolak (1994) derives and implements a procedure to recover a consistent estimate of a regulated firm's production technology taking into account the impact of private information on regulator–utility interactions. As noted above, this task requires the imposition of parametric and distributional assumptions. These assumptions allow Wolak to identify the underlying economic primitives from the joint density of the regulated price, the firm's output, input choices and total cost, conditional on the vectors of input prices and demand shifters, $h(p_Q, Q, X, C \mid p_X, Z)$. As we have said repeatedly, there is no single "right" way to make these assumptions. It is presumably economics and the specific features of a market and regulatory environment that can help the researcher defend the assumptions necessary to obtain identification.

Wolak models the behavior of a sample of Class A California Water utilities using annual data on utility outputs, production costs, input quantities and prices, several demand shifters, and output prices. He has panel data from 1980 to 1986. Class A utilities distribute water and provide services to large cities in California. Consistent with our earlier discussion, the California Public Utilities Commission (CPUC) sets the retail price of water for these utilities on a prospective basis.

As Wolak (1994) notes, the water supply industry was chosen, as opposed to other regulated industries, such as telecommunications, or electricity, for two major reasons. First, the structure of production in water delivery is extremely simple relative to producing electricity or providing telecommunications services. Second, the assumption of a single homogenous product is likely to be far less objectionable than would be the case for either telecommunications or electricity. These reasons help Wolak to simplify his econometric model.

As with any structural econometric modeling exercise, it is important to have a clear idea of what economic magnitudes can be recovered from a structural model. Wolak would first like to obtain a consistent estimate of the underlying production function. To do this, he explicitly models the impact of the utility's private information on production. Instead of estimating the production function directly, Wolak derives the utility's cost function under the assumption of expected profit-maximizing behavior. He then estimates the production function parameters from the cost function. A useful by-product of this approach is an estimate of the distribution of private information. A second goal

of the structural model is to obtain an estimate of how much firm output is distorted from minimum cost production due to the presence of private information. A third goal is to test the relative performance of the asymmetric information model versus the conventional symmetric information model of the regulator–utility interaction.

To evaluate the relative performance of the asymmetric information model, the paper posits a second behavioral model of regulator–utility interaction for the same set of underlying economic primitives. In this model, the utility initially possesses private information. Through information gathering, however, the regulator is able to completely learn this parameter. Consequently, the regulator can impose what Wolak calls the symmetric (or full) information regulatory outcome. Unfortunately, the econometrician is unable to observe this private information parameter and so must take it into account.

Wolak does this when specifying and estimating the behavioral cost functions. The asymmetric information model assumes that the utility possesses private information, but the regulator is unable to completely learn this private information through its information gathering efforts. However, the regulator does learn the distribution of this private information for each utility, and regulates using this incomplete information optimally. The regulator is assumed to impose a version of the asymmetric information optimal "second-best" regulatory outcome described in the previous section. In this case, the econometrician also is unable to observe the utility's private information (or even its distribution), but must account for this assumed utility–regulator interaction when estimating the parameters of the utility's production function.

Wolak assumes the production function for water delivery for utility i is

$$Q_i = f\left(K_i, L_i^*, E_i, \epsilon_i^Q \mid \beta\right), \tag{122}$$

where K_i denotes capital (physical plant and water sources), L_i^* labor, and E_i electricity. The parameter β is a vector describing the technical coefficients of production. It is known to both the regulator and utility, but is unknown to the econometrician. The variable ϵ_i^Q is a stochastic disturbance to the ith utility's production process that is realized after the utility makes its capital stock selection, but before it produces. The utility knows the distribution of ϵ_i^Q, which is independently and identically distributed over time and across utilities. Allowing for this source of unobservable heterogeneity in the production function increases the realism of the model because there are a number of factors that are unknown to the firm at the time it chooses the configuration and capacity of its water distribution network. (A utility's distribution network is a major component of its capital stock.)

To account for all these forms of unobserved heterogeneity, Wolak must make parametric and distributional assumptions to identify the underlying economic primitives from $h(p_Q, Q, X, C \mid p_X, Z)$. Without these assumptions, it is impossible to proceed. Once again, this illustrates our point that it is specific parametric economic and statistical assumptions that allow us to go from the statistical joint distribution of the data, $h(p_Q, Q, X, C \mid p_X, Z)$, to statements about the production technologies, market demand and information primitives.

The source of the utility's private information is the efficiency of its labor input. To this end, Wolak makes the distinction between, L_i^*, the amount of labor actually used in the production process, and L_i, the observed physical quantity of labor input which is implied by the utility's total labor costs. These two magnitudes are related by the equation $L_i^* = L_i/\mathrm{d}(\theta_i)$, where $\mathrm{d}(\theta)$ is a known increasing function and θ_i is interpreted as utility i's labor inefficiency parameter. (Higher values of θ_i imply more inefficiency.) The econometrician and regulator observe the utility using the quantity of labor L_i, but the actual amount of "standardized" labor available in the production process is L_i^*. This specification is based on the fact that labor costs are a major component of total maintenance expenditures, and system maintenance is a major determinant of water system efficiency. Thus, while the regulator can observe how much labor is employed at the utility, L_i, it does not know the productivity of this labor. The utility's *ex post* observed costs have the form $w_i L_i + r_i K_i + pe_i E_i$, where w_i is the wage rate, r_i is the price of capital, and pe_i is the price of electricity. Note that the utility pays for observed labor, L_i.

From the viewpoint of the econometrician, θ_i is an unobservable random variable that determines the productivity of labor. In this sense, it is comparable to other unobservables, such as ϵ_i^Q. What is special about θ_i as an unobservable is that it may also be unobserved by the regulator. This is the case of Model A, the asymmetric information model. There, the regulator only knows the distribution of θ_i, $F(\theta)$, and thus can only condition its decisions on that information – much like what happens in an auction. By contrast, in the symmetric information model (Model S), the θ_i plays the role of unobserved heterogeneity because the regulator and firm observe it but the econometrician does not.

For both Model S and Model A, the utility chooses its input mix to maximize expected profits given its private information. Each utility faces the demand function $Q^D = Q_i(p_i)\epsilon_i^D$ for its product, where ϵ_i^D is a positive, mean one stochastic shock to demand. This shock is assumed independently and identically distributed across time and utilities. Once p is set, the demand shock is realized; the utility then produces output to satisfy demand (which in both models is known both to the regulator and the utility).

Because the utility's price and capital stock are set before the utility produces each period, the utility's desire to maximize expected profits will lead it to minimize total operating costs under both Models A and S for a fixed level of output and capital stock. Thus, for each model, Wolak can compute a conditional variable cost function $\mathrm{CVC}(pe, w, \theta, K, Q, \epsilon^Q, \eta_L, \eta_E \mid \beta)$, where η_L and η_E are mean one optimization errors. Wolak introduces these errors to allow for the fact that the first-order conditions for L and E do not hold exactly. Note that the utility's private information, θ, enters into the conditional variable cost function.

Using this expression for variable costs, utility i's total observed costs equal:

$$\mathrm{TC} = \mathrm{CVC}(pe, w, \theta, K, Q, \epsilon_q, \eta_L, \eta_E \mid \beta) + r_i K_i. \tag{123}$$

As noted earlier, the firm's capital stock serves two roles: (1) it reduces the total cost of serving demand; and, (2) it signals to the regulator the firm's true productive efficiency. This tension between increasing profits by choosing the minimum total cost level of capital and increasing the size of the capital stock in an effort to be rewarded by the regulator with a higher output price, leads to distortions from least-cost production by the firm.

9.3.1. Symmetric information model

In the symmetric information model, the regulator observes each utility's true θ and sets the monthly fixed fee (F) and per unit price (p) to maximize expected consumer surplus subject to the constraint that the utility's expected profits (with respect to the distributions of ϵ^Q and ϵ^D) equal zero. This implies that the regulator will solve for the p, T, and K which maximize expected consumer surplus for the utility's consumers.

Let $S_i(p) = E_D(\epsilon_i^D)\int_p^\infty Q_i(s)\,ds$ denote expected consumer surplus for the ith utility, where $E_D(.)$ denotes the expectation with respect to the distribution of ϵ^D. In terms of our notation, the regulator solves:

$$\begin{aligned}
&\max_{p,T,K} S_i\big[p(\theta_i)\big] - T(\theta_i)\\
&\text{subject to}\quad E_{QD}\big(\pi(\theta_i)\big) = E_{Qd}\big[p(\theta_i)Q\big[p(\theta_i)\big]\epsilon_i^D + T(\theta_i)\\
&\qquad\qquad - \mathrm{CVC}\big(pe, w, \theta_i, K(\theta_i), Q(\theta_i)\epsilon_i^D, \epsilon_i^Q, \eta_i \mid \beta\big)\big]\\
&\qquad\qquad - r_i K(\theta_i) = 0, \qquad (124)
\end{aligned}$$

where $E_{Qd}(.)$ is the expectation with respect to the distribution of both ϵ^Q and ϵ^D and $\eta_i = (\eta_i^L, \eta_i^E)'$ is the vector of optimization errors from the conditional variable cost function optimization problem. The first-order conditions for the regulator's problem imply:

$$p_i = \frac{\partial E_{QD}[\mathrm{CVC}(pe, w, \theta_i, K(\theta_i), Q(\theta_i), \epsilon_i^D, \epsilon_i^Q, \eta_i \mid \beta)]}{\partial Q}, \qquad (125)$$

$$r_i = -\frac{\partial E_{QD}[\mathrm{CVC}(pe, w, \theta_i, K(\theta_i), Q(\theta_i), \epsilon_i^D, \epsilon_i^Q, \eta_i \mid \beta)]}{\partial K}. \qquad (126)$$

The fixed fee, $T(\theta_i)$, is set so that expected profits are zero at the values of $K(\theta_i)$ and $p(\theta_i)$ that solve (125) and (126).

9.3.2. Asymmetric information model

In the asymmetric information model (Model A), the regulator recognizes that the utility may mis-report θ as higher than it really is (i.e., the utility claims to be less efficient than it really is). Consequently, the regulator constructs price, fixed fee and capital stock (as a function of θ) such that given these schedules, the utility finds it profit-maximizing to

report its true θ. The regulator picks price, the fixed fee, and the capital stock that maximize expected (with respect to the distributions of θ, ϵ^D, and ϵ^Q) consumer surplus.

To derive the Model A equilibrium, Wolak follows the approach given in Baron (1989). The first step is to determine the global truth-telling constraints. A utility with true parameter θ_x that reports θ_y earns expected profit

$$E_{QD}\big[\pi(\theta_y,\theta_x)\big] = E_{QD}\big[p(\theta_y)Q\big(p(\theta_y)\big)\epsilon^D - \mathrm{CVC}\big(\theta_x, K(\theta_y), Q(\theta_y)\big)\big] - rK(\theta_y) + T(\theta_y), \tag{127}$$

where we suppress the dependence of the minimum variable cost function (CVC) on pe, w, ϵ^Q and ϵ^D, η and β. Consequently, for any two arbitrary values θ might take for a given utility, say θ_x and θ_y, incentive compatibility requires $E_{QD}[\pi(\theta_x,\theta_x)] \geqslant E_{QD}[\pi(\theta_y,\theta_x)]$, meaning that the firm expects to earn higher profits by announcing θ_x when its true type is θ_x, than it expects to earn from announcing any other $\theta_y \neq \theta_x$. The next step is to specify the local version of this global constraint:

$$\frac{\mathrm{d}E_{QD}[\pi(\theta)]}{\mathrm{d}\theta} = -\frac{\partial E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial\theta}, \tag{128}$$

for all $\theta \in [\theta_l, \theta_h]$. Equation (128) is the local incentive compatibility condition that quantifies how rapidly the regulator must raise the expected profits of a utility as its true θ value falls (the utility becomes more efficient) in order to encourage truthful revelation. By integrating (128), one obtains the expected profit function. This implies that the expected profit function is locally decreasing in θ so that the participation constraint, which requires the firm to earn nonnegative expected profits for all values of θ, can be replaced by the single constraint that $E_{QD}[\pi(\theta_h)] \geqslant 0$, where θ lies in the interval $[\theta_l, \theta_h]$.

The regulator's optimization problem is

$$\begin{aligned}
&\max_{p(\theta),T(\theta),K(\theta)} \int_{\theta_l}^{\theta_h} \big[S_i\big(p(\theta)\big) - T(\theta)\big] f(\theta)\,\mathrm{d}\theta \\
&\text{subject to}\quad E_{QD}\big(\pi(\theta_i)\big) = E_{QD}\big[p(\theta_i)Q\big(p(\theta_i)\big)\epsilon_i^D + T(\theta_i) \\
&\qquad\qquad - \mathrm{CVC}\big(pe, w, \theta_i, K(\theta_i), Q(\theta_i), \epsilon_i^D, \epsilon_i^Q, \eta_i \mid \beta\big)\big] \\
&\qquad\qquad - r_i K(\theta_i) \\
&\qquad \frac{\mathrm{d}E_{QD}[\pi(\theta)]}{\mathrm{d}\theta} = -\frac{\partial E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial\theta}, \\
&\qquad E_{QD}\big[\pi(\theta_h)\big] \geqslant 0.
\end{aligned} \tag{129}$$

Although Wolak does not explicitly include the restrictions implied by the global truth-telling constraints, he derives restrictions on the regulatory environment and distribution of θ necessary for the price, capital, and fixed fee functions that solve (129) to also satisfy the global incentive compatibility constraints.

Note that the formulation in (129) refers specifically to the ith utility–regulator pair. Because the regulator does not know utility i's efficiency parameter, she must set p, T,

and K functions over the entire support of θ for each utility. Consequently, the regulator must solve this problem for each utility that it regulates.

Wolak (1994) presents a detailed discussion of the derivation of the first-order conditions for this optimization problem. For our purposes, we simply want to show how these first-order conditions differ from those for the Model S solution. The first-order conditions analogous to (125) and (126) for Model A are:

$$p(\theta) = \left[\frac{\partial E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial Q} + \frac{F(\theta)}{f(\theta)}\frac{\partial^2 E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial\theta\partial Q}\right]\eta_p, \tag{130}$$

$$r = -\left[\frac{\partial E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial K} + \frac{F(\theta)}{f(\theta)}\frac{\partial^2 E_{QD}[\mathrm{CVC}(\theta, K(\theta), Q(\theta))]}{\partial\theta\partial K}\right]\eta_K, \tag{131}$$

where η_p and η_K are the mean one multiplicative optimization errors added for same reasons given above in the discussion of the Model S solution. These two equations determine the amount of capital stock $K(\theta)$ a utility of type θ will purchase, and the price $p(\theta)$ it will be directed to charge. The demand function $Q_i(p)$ and these two equations determine the two regulatory variables $K(\theta)$ and $p(\theta)$. The fixed fee $T(\theta)$ is given by

$$T(\theta^*) = E_{QD}\big[\pi(\theta^*)\big] - E_{QD}\big[p(\theta^*)Q\big(p(\theta^*)\big)\epsilon^D + \mathrm{CVC}\big(\theta^*, K(\theta^*), Q(\theta^*)\big)\big] + rK(\theta^*) \tag{132}$$

for a utility of type θ^*. Once a utility's K is chosen and its p and T are set, its demands for L and E can be determined from the solution to the minimum operating cost problem.

These first-order conditions demonstrate that the presence of asymmetric information in the regulator–utility interaction leads to both deviations from minimum cost production and efficient output prices in the sense that price differs from marginal cost. As discussed above, this deviation from minimum cost production occurs because the firm also uses its capital stock to signal to the regulator its greater productive efficiency and therefore lower value of θ.

9.4. Econometric model

Following our procedure outlined in Section 4 for constructing structural econometric models, this section discusses the functional form of the production function $Q_i = f(K_i, L_i^*, E_i, \epsilon^Q \mid \beta)$ and derives the cost function which is used to recover an estimate of the parameter vector β. We then discuss the specification of distributions for the structural disturbances introduced into the model and derive the likelihood function. Wolak's model contains the first three types of disturbances discussed in Section 4: (1) unobserved heterogeneity in the form of the utility's private information θ_i, (2) shocks which agents in the model optimize against (ϵ^Q and ϵ^D), (3) optimization errors which allow

agents' first-order conditions to only be satisfied in expectation (η_j, $j = L, E, K, p$). Appendix A of Wolak (1994) shows that the composite errors to the structural equations are functions of these disturbances.

Wolak's choice of fairly simple functional forms for the production function and demand function allows him to impose conditions on the parameters of the underlying econometric model that guarantee a solution to the regulator's problem. More flexible functional forms for $Q_i = f(K_i, L_i^*, E_i, \epsilon^Q \mid \beta)$ would not allow this. These functional forms allow constraints on the parameters of the economic environment which guarantee the existence of a Model A solution. These functional forms allow Wolak to perform counterfactual experiments with his parameter estimates which illustrate several important empirical distinctions among Model S, Model A, and conventional estimation procedures.

Wolak uses the Cobb–Douglas production function $Q = \beta_0 K^{\beta_K} (L/\mathrm{d}(\theta))^{\beta_L} E^{\beta_E} \epsilon^Q$, where $\mathrm{d}(\theta) = \theta^{(\beta_L+\beta_E)/\beta_L}$. The demand function for the utility's output is

$$Q_d = \begin{cases} \exp(Z'b) p^{-\kappa} \epsilon^D & \text{if } p \leqslant p_{\max}, \\ 0 & \text{if } p > p_{\max}, \end{cases} \tag{133}$$

where Z is a vector of utility service area characteristics assumed shift demand, b is a parameter vector associated with Z, κ is the elasticity demand for water, and $p_{\max}$ is the price beyond which demand for the firm's output is zero.

Solving the minimum operating cost problem for this production function yields the following (conditional on K) variable cost function:

$$\begin{aligned} &\mathrm{CVC}(pe, w, K, Q, \theta, \epsilon \mid \beta) \\ &\quad = \theta \beta_0^{-\frac{1}{\beta_L+\beta_E}} K^{-\frac{\beta_K}{\beta_L+\beta_E}} \left[\left(\frac{\beta_L}{\beta_E} \right)^{\frac{\beta_E}{\beta_L+\beta_E}} + \left(\frac{\beta_L}{\beta_E} \right)^{-\frac{\beta_L}{\beta_L+\beta_E}} \right] \\ &\qquad \times w^{\frac{\beta_L}{\beta_L+\beta_E}} pe^{\frac{\beta_E}{\beta_L+\beta_E}} Q^{\frac{1}{\beta_L+\beta_E}} u, \end{aligned} \tag{134}$$

where u is a function of our previously defined disturbances ϵ^D, η_L and η_E and the parameter vector β.

Taking the partial derivative of the expected value of this cost variable function, $E_{QD}[\mathrm{CVC}]$, with respect to K and inserting it into the first-order condition for the symmetric information regulatory outcome with respect to K, yields the following unconditional variable cost (VC) function:

$$\mathrm{VC}(S) = D^* r^\alpha w^\gamma \theta^{(1-\alpha)} pe^{(1-\alpha-\gamma)} Q_d^\delta \nu. \tag{135}$$

Expressions for D^* and ν in terms of the underlying parameters of the model are given in Appendix A of Wolak (1994). The parameters α, γ and δ are defined as follows:

$$\alpha = \frac{\beta_K}{\beta_K + \beta_L + \beta_E}, \qquad \gamma = \frac{\beta_L}{\beta_K + \beta_L + \beta_E}, \tag{136}$$

$$\delta = \frac{1}{\beta_K + \beta_L + \beta_E}. \tag{137}$$

The only difference between this unconditional variable cost function and the usual Cobb–Douglas unconditional variable cost function is the presence of the utility's private information, θ.

We should emphasize that because it excludes capital costs, this is the utility's minimum variable cost function conditional on θ – not the minimum total cost function. Although it is straightforward to derive the utility's minimum total cost function from (125), Wolak departs from the tradition of estimating a total cost function for the following reason. Operating or variable costs are measured with little if any error, whereas, capital cost (the missing ingredient necessary to compute total production costs) is extremely poorly measured. Rather than complicate the analysis with potentially substantial measurement error, he instead uses the variable cost function to estimate the same parameters of the utility's production function that can be recovered by estimating a total cost function.

To derive the asymmetric information cost function, substitute the partial derivative of the expected value of the variable cost function, $E_{QD}(\mathrm{CVC})$, with respect to K into the first-order condition for the optimal capital stock given in (130). Simplifying this expression gives the following variable cost function:

$$\mathrm{VC}(A) = D^* H(\theta)^{-\alpha} \theta r^{\alpha} w^{\gamma} pe^{(1-\alpha-\gamma)} Q_d^{\delta} \nu, \tag{138}$$

where $H(\theta) = [\theta + \frac{F(\theta)}{f(\theta)}]$. The parameters α, γ and δ are as defined above.

The final step toward developing the structural econometric model is to specify distributions for all of the stochastic shocks to the econometric model. This step is needed to derive the likelihood function for the variable cost functions under the two information structures. Wolak requires that ν be lognormally distributed with $\ln(\nu) \sim N(\mu_\nu, \sigma_\nu^2)$ independent across time and utilities.

Taking the natural logarithm of both sides of (135) gives the following symmetric information logarithm-of-variable-costs equation:

$$\begin{aligned}\ln\big(\mathrm{VC}(S)\big) = {} & \xi^* + (1-\alpha)\ln(\theta) + \gamma \ln(w) + \alpha \ln(r) + (1-\alpha-\gamma)\ln(pe) \\ & + \delta \ln(Q_d) + \zeta,\end{aligned} \tag{139}$$

where $\xi^* = \ln(D^*) + \mu_\nu$ and $\zeta = \ln(\nu) - \mu_\nu$. Therefore, ζ is $N(0, \sigma_\zeta^2)$, where $\sigma_\zeta^2 = \sigma_\nu^2$. Repeating this procedure for Equation (138) yields the asymmetric information log-of-variable-costs equation:

$$\begin{aligned}\ln\big(\mathrm{VC}(A)\big) = {} & \xi^* - \alpha \ln\big(H(\theta)\big) + \gamma \ln(w) + \alpha \ln(r) + (1-\alpha-\gamma)\ln(pe) \\ & + \delta \ln(Q_d) + \zeta.\end{aligned} \tag{140}$$

The final step of the process is to define the likelihood function for each information structure. First we define notation which simplifies the presentation. Let $\Gamma^* = (\xi^*, \alpha, \gamma, \delta)'$. Define $X = (\ln(r), \ln(w), \ln(pe))'$, $q = \ln(Q_d)$, and $Y = \ln(VC)$. In this notation we can abbreviate Equations (139) and (140) as:

$$Y = \Omega_Y(X, q, \Gamma^*, \theta) + \zeta, \tag{141}$$

$$Y = \Psi_Y(X, q, \Gamma^*, \theta) + \zeta, \tag{142}$$

where $\Omega_Y(X, q, \Gamma^*, \theta)$ is the right-hand side of (139) excluding ζ and $\Psi_Y(X, q, \Gamma^*, \theta)$ is the right-hand side of (140) excluding ζ.

We now derive the likelihood function and discuss the estimation procedure for the case of Model S. Following this discussion, we describe the additional complications introduced by Model A. Under Wolak's assumptions on the functional form for the production function and the aggregate demand function the equilibrium value of q under Model S is

$$q = \big(Z', X', \ln(\theta)\big)\Lambda^* + \psi, \tag{143}$$

where Λ^* is the vector of coefficients associated with $(Z', X', \ln(\theta))$ and ψ is assumed to be joint normally distributed with ζ. Let $\rho_{\zeta,\psi}$ denote the correlation between ζ and ψ. Finally, define $\Lambda = (\Lambda^{*\prime}, \sigma_\psi^2, \rho_{\zeta,\psi})'$. Conditional on the value of θ, Equations (141) and (143) make up a triangular system of simultaneous equations. The determinant of the Jacobian of the transformation from $(\zeta, \psi)'$ to $(Y, q)'$ is one, so that the joint density of $(Y, q)'$ conditional on θ, X and Z is

$$\begin{aligned} &h_S\big(Y, q \mid \ln(\theta), \Gamma, \Lambda\big) \\ &\quad = \frac{1}{2\pi\sigma_\zeta^2\sigma_\psi^2(1-\rho_{\zeta,\psi}^2)^{1/2}} \\ &\qquad \times \exp\left[-\frac{1}{2(1-\rho_{\zeta,\psi}^2)}\big[(\psi/\sigma_\psi)^2\big] - 2\rho_{\zeta,\psi}(\psi\zeta)/(\sigma_\psi\sigma_\zeta) + (\zeta/\sigma_\zeta)^2\right], \end{aligned} \tag{144}$$

where $\Gamma = (\Gamma^*, \sigma_\zeta)'$. Note that θ enters both (141) and (143) only through $\ln(\theta)$, so that without loss of generality we can express $h_S(\cdot,\cdot)$ as a function of $\ln(\theta)$. Because θ is unobservable, to construct the likelihood function in terms of the observable variables, we must compute the density of (Y, q) given X and Z only. To obtain this density we integrate the conditional density $h_S(Y, q \mid \ln(\theta), \Gamma, \Lambda)$ with respect to the density of θ. Integrating with respect to the density of θ, yields

$$g\big(Y, q \mid X, Z, \Gamma, \lambda, F(.)\big) = \int_{\theta_l}^{\theta_h} h_S\big(Y, q \mid X, Z, \ln(\theta), \Gamma\big) f(\theta)\,\mathrm{d}(\theta). \tag{145}$$

This likelihood function is similar to Porter's regime switching model likelihood function. In Porter's case I_t is the unobserved regime indicator and in the present case θ is a continuously distributed random variable with compact support. In the same way that Porter was able to identify the density of I_t from his assumption of conditional normality of the density of equilibrium price and quantity, Wolak (1994) is able to identify the distribution of θ from the joint normality assumptions of Y and q. In addition, in the same sense that the economic structure of competitive and collusive pricing regimes was identified by the conditional normality assumption in Porter's model, the primitives of the private information regulator–utility interaction are identified by the conditional normality assumption in Wolak's model.

The construction of the likelihood function for the asymmetric information case proceeds in an analogous fashion, with the major complication being the presence of $H(\theta)$, which is a function of both $f(\theta)$ and $F(\theta)$ in both regression equations. The conditional density of $(Y, q)'$ given θ, X and Z under Model A takes the same form as for Model S with Equation (141) replaced by Equation (142) and the log-output Equation (143) replaced by the following equation:

$$q = \big(X', Z', \ln(\theta), \ln\big(H(\theta)\big)\big)\Phi + \psi, \tag{146}$$

where Φ is the vector of coefficients associated with $(X', Z', \ln(\theta), \ln(H(\theta))')$.

The conditional distribution of Y and q given Z, X, and θ for this information structure, $h_A(Y, q \mid X, Z, \theta)$, depends on θ through both $\ln(\theta)$ and $\ln(H(\theta))$. To construct the likelihood in terms of only observables, we integrate this conditional density with respect to $f(\theta)$ over the interval $[\theta_l, \theta_h]$.

For both Model S and Model A, conventional maximum likelihood estimation procedures can be applied to compute the coefficient estimates and their standard errors.

9.5. Estimation results

A major goal of the empirical analysis is to recover characteristics of production process, and in particular, the returns to scale in production, accounting for the impact of the utility's private information. Wolak finds that applying conventional minimum cost function Cobb–Douglas estimation techniques, the returns to scale estimates obtained that are implausibly high, with cost elasticities with respect to output estimates as high as 0.77, which means that a 10 percent increase in output only increases total costs by 7.7%. Other estimates were even lower. However, applying the maximum likelihood estimation techniques outlined above for the Model S and Model A solutions, Wolak finds cost elasticities with respect to output greater than 1, for both the Model S and Model A estimates, which implies slight decreasing returns to scale in production, although the null hypothesis of constant returns to scale cannot be rejected. This dramatic difference in returns to scale estimates points out the importance of controlling for this unobserved firm-level heterogeneity in productive efficiency when attempting to recover consistent estimates of the characteristics of the regulated firm's production process.

Wolak is also able to recover estimates of $F(\theta)$, which determines the form of the optimal regulatory contract under asymmetric information. Armed with this information he is able to compute the following counterfactuals for each point in his dataset using the Model A parameter estimates. First, he computes the ratio of total operating costs under a Model A solution versus the Model S solution holding constant the level of output produced by the firm under both scenarios. This answers the question of how much less costly, in terms of variable costs, it is to produce a given level of output under the Model A versus Model S versions of the regulatory process. Wolak also performs this same counterfactual for total production costs and finds that in terms of total production costs, the same level of output costs approximately 5–10 percent more to provide

under the Model A regulatory process relative to the Model S regulatory process. These distortions from minimum cost production occur because the more efficient firms find it profitable to signal their superior productive efficiency to the regulator through their capital stock choice.

Wolak also computes the welfare cost to consumers from asymmetric information by comparing the market-clearing level of output under the Model A solution versus the Model S solution for the same values of the input prices and θ. He finds the output level produced under the Model A solution is roughly 20 percent less than the level of output under the Model S solution for the Model A parameter estimates, which indicates a significant welfare loss to consumers associated with asymmetric information.

In an attempt to see whether Model A or Model S provides a statistically superior description of the observed data, Wolak performs a nonnested hypothesis test of Model A versus Model S. He finds that Model A provides a statistically significantly superior description of the observed data relative to Model S. As discussed in Section 3, this does not validate Model A as the true model for the regulatory process. It only states that for the same functional forms and economic primitives, the strategic interaction implied by Model A provides a statistically superior description of the observed data.

9.6. Further extensions

There are variety of directions for future research in this area given the enormous number of competing theoretical models of the private information regulator–utility interaction. Sorting through the empirical implications of these models across a variety of regulated industries would help to focus future theoretical and empirical research in this area. Recent work in this area includes: Dalen and Gomez-Lobo (1997) who study the impact of these incentive contracts in the Norwegian Bus Transport Industry and Gagnepain and Ivaldi (2002) who assess the impact of incentive regulatory policies for public transit systems in France.

10. Market structure and firm turnover

So far we have discussed IO models in which the number of market participants (e.g., firms or bidders) is given. IO economists have recently devoted considerable energy toward modeling how changes in market structure can affect the extent of competition in a market. In particular, the theoretical literature has explored two related questions:

1. "How many competitors are needed to insure effective competition?" and
2. "What factors encourage firms to enter markets?"

Theoretical answers to these questions often hinge delicately on the assumptions made about firms' costs, market demand and firms' conjectures about competitors' behavior. Unfortunately, there are very few structural econometric models that would allow one to identify the empirical relevance of demand, cost and strategic explanations. In large part

this is because competition models in which the number of participants is endogenous are complicated and difficult to solve.

Only recently have empirical researchers begun to make progress in developing structural econometric models that can speak to specific strategic models of entry and entry deterrence. In this section we outline some of the econometric issues associated with modeling the number of firms in oligopolistic markets. Again, our intent is not so much to survey the literature as to show what one can learn from information about the number and identities of firms in a market.[32] We shall see that while structural models of entry, exit and market structure raise many of the modeling issues discussed in Sections 5–9, there are also new issues.

10.1. Overview of the issues

Sections 5–7 showed how economists have used information about the joint density of prices and quantities $f(P, Q \mid X, Z) = f(P_1, \ldots, P_N, Q_1, \ldots, Q_N \mid X, Z)$ to recover information about firms' demand curves and costs. In general, the conditional density $f(\cdot)$ is a statistical object, and a high-dimensional one at that. In practice this means that it would be hopeless to try and estimate a $2 \times N$ conditional joint density nonparametrically from market-level data. While going to consumer-level data can improve inferences, in general it will be extremely difficult to obtain the representative consumer-level datasets necessary to estimate flexible and yet precise estimates of firm-level demands. These observations suggest that considerable economic structure will have to be introduced if one is to obtain meaningful estimates of firms' demands and costs.

The literatures discussed in Sections 5–9 presume that the number of firms is exogenous. One consequence of this assumption is that N enters objects such as $f(P, Q \mid X, Z)$ as a conditioning variable rather than something to be explained. One way to make N endogenous is to imagine that each market has the same $M > N$ potential entrants. Each of these potential entrants makes a discrete decision whether or not to enter. The conditional density of the market data and these entry decisions is $f(P_1, \ldots, P_M, Q_1, \ldots, Q_M, a_1, \ldots, a_M \mid X, Z, W, M)$. Here, the a_i are zero-one indicators for whether or not potential entrant i has entered and W are any new conditioning variables.

This expression makes it easy to appreciate why many studies do not make N endogenous. First, there are many different collections of the a_i that yield the same N. In principle, the researcher might wish to explain not just N but why a particular ordering of the a_i was obtained. Second, because the dimensionality of $f(\cdot)$ has gone up considerably, it becomes even more difficult to estimate nonparametrically. For example, it seems unlikely that a researcher would have a large sample of markets that have the same number of potential entrants M. Finally, the form of $f(\cdot)$ may differ with the identities of each entrant.

[32] For a more complete discussion see Berry and Reiss (in press).

Because nonparametric methods are impractical, the researcher will have to impose economic structure to get anywhere. In particular, now the researcher will have to add equations that explain each of the a_i. These conditions must explain why some but not other potential entrants entered the market.

Our discussion so far has dealt with more obvious complications introduced by making N and the identities of entrants endogenous. There are less obvious complications as well. Two of the most critical are that: (1) the underlying theory may deliver ambiguous predictions about which firms will enter in equilibrium; and, (2) the underlying theory may deliver no (pure-strategy) predictions about which firms will enter in equilibrium. These are new complexities, ones we did not really see in Sections 5–9. Before we explore their significance for structural modeling, it is useful to back up and provide a broader sense of the types of economic issues that one might hope to address with structural models of market concentration and competition.

10.1.1. Airline competition and entry

Since the deregulation of US passenger airline markets in the late 1970s, travelers and economists have speculated about whether sufficient competition exists in different city-pair markets.[33] One does not have to look far to understand why. Travelers routinely encounter wide disparities in an airline's fares (per seat mile) over time, across routes and even for seats on the same flight. Despite this considerable variation in a given airline's fares, there appears to be much less variation in fares across competing carriers. Industry critics contend that such patterns are obvious evidence of ineffective competition. They also argue that high concentration on some individual city-pair routes contributes to the problem. Some industry advocates argue the opposite. They contend that fare matching is evidence of competition, and that fare differences at worst reflect price discrimination. Some also claim that high concentration is evidence of economies of scale and route density, and that entry (or the threat of entry) of small upstart carriers is enough to insure effective competition.

These two views provide a challenge to IO economists, and there have been many attempts to distinguish between them. To delve deeper, it is useful to imagine that we have data (consistent with the US experience) indicating that short haul routes between small cities tend to be highly concentrated and have high (per seat mile) fares. The technological and demand explanation for this correlation is that the costs of service on these routes is high relative to demand. Thus, some routes will have so little demand relative to costs, that at most one firm can profitably serve the market. This one firm would behave as a monopolist and charge high prices to recover its costs. The anti-competitive explanation for the observed correlation is that high concentration and fares are the result of strategic behavior. For example, even if the small market could support

[33] See for example Borenstein (1992), Brueckner, Dryer and Spiller (1992), Morrison and Winston (1996), Ott (1990), and Windle (1993).

many carriers, dominant carriers can convince potential entrants that entry would be met with stiff competition.

Can we distinguish between these explanations? Our answer is: given the current state of the theory, econometric models and data, we cannot generally. The main reason is that much of what the theory points us toward is unobservable. Researchers do not observe the marginal and fixed costs that are central to technological explanations. We also do not observe potential entrants' expectations about incumbent behavior, which are central to strategic explanations. Does this mean we cannot learn anything from a structural model of market structure? The answer to this is no.

What we can imagine doing in principle is building structural models that would examine how alternative competitive models fit the data. For instance, we might begin in the spirit of the models in Sections 5–7 by writing down functional forms for city-pair demand, and firms' fixed and variable costs. This is not, however, as easy as it sounds. Prior studies have documented that airlines' costs of service depend in complex ways not only on route-specific factors, such as miles traveled, airport fees, etc., but also on network and fleet characteristics (e.g., whether the plane will carry passengers beyond a city or transfer passengers at a hub and code-sharing agreements). Nevertheless, we might attempt a parametric model of demand and costs. At that point, unlike most of the models in Sections 5–7, we would have to grapple with the problem that the number of carriers in a market is endogenous: it is affected by demand and supply conditions. We therefore also have to model how fixed and marginal costs impact the number of firms in the market (and possibly the identities of those firms).

Here, we encounter tricky specification issues. Economic theory suggests that to model the number of firms we need to model why (and possibly which) firms did not enter. But this involves modeling potential entrants' expectations about what would happen after entry, something we never observe. Moreover, because the same carriers compete with each other in other markets, we may have to model how actions in any one market affect outcomes in other markets.

At this point, it might seem that a complete structural model of airline competition is hopeless. There is, however, something that we can learn with the right data. The critical events that tell us something about competition and market structure are instances of entry and exit. Consider, for example, our sample of small markets. In principle, we observe some city-pair markets in which there is no (direct) service, others in which there is a monopoly, a duopoly, and so on. If (and this is an important if) we can control for factors that might lead to cost of service and demand differences across markets, then we can ask how much demand does it take to support at least one carrier. This level of demand tells us something about a single carrier's fixed and marginal costs relative to demand. We can then compare this level of demand to what it takes to support a second firm in the market. This level of demand tells us more about costs and potentially behavior. Suppose, for instance, we do not observe a second carrier enter a city-pair market until demand is roughly twenty times what it takes to support a single carrier. One's intuition is that if the second carrier has the same costs and product as the first,

that this difference must reflect pessimism on the part of the second carrier as to value of entering a monopoly market.

It is this type of intuition that structural models of the number of firms, or entry and exit seek to make more precise. That is, the goal of a structural model is to show how changes in population and other exogenous market conditions affect the (apparent) ability of potential entrants to cover costs. The primary value of a formal model is that it makes clear what economic and stochastic assumptions are necessary, given the available data, to isolate differences between firms' costs and the expectations they have about post-entry competition.

10.2. An economic model and data

Our airline example makes three points that are worth emphasizing. First, debates about the competitiveness of markets often hinge on assumptions about what determines a market's structure (e.g., the number of firms). Second, some of the most critical factors affecting the ease of entry and exit are unobservable (e.g., firms' fixed and marginal costs, and expectations about post-entry competition). Third, while we can potentially use structural models to draw inferences about the unobservables present in IO theories, these models, like all structural models, will contain untestable assumptions. These assumptions may be too numerous to be credible.

An important corollary to this third point is that the form of the data available will have an important impact on what we can estimate. In our airline example, for instance, we might have data on a cross section of similar city-pair markets or time series data on the same market over time. Both of these data sets raise modeling issues. In cross-section data we have to worry about changes in the identity and number of potential entrants across markets. We may also have to worry that the behavior of firms in one market may affect their behavior in other markets. While time-series data have the advantage of holding constant market-specific conditions, researchers must again worry that the firms' decisions may be linked through time. When they are, it makes sense to model firms' decisions using dynamic games. While some progress has been made in formulating and solving such games, to date their computational demands have largely made them impractical for empirical work. As a consequence, almost all structural market structure models are static.

Most empirical work in this area has tended to rely on cross-section data. As such they focus on modeling which firms are producing, as opposed to firm turnover; i.e., which firms are entering or exiting. In a typical cross-section application, a researcher might have data on

1. the number of potential entrants into each market, M;
2. the entry decisions of each potential entrant: $a = (a_1, a_2, \ldots, a_N)$;
3. market-specific information X (e.g., market size); and
4. firm-specific information, $Z = (z_1, z_2, \ldots, z_M)$ (e.g., identities and product characteristics).

In addition, in an ideal application the researcher may also observe the prices and quantities of actual entrants: $P_1, \ldots, P_N$ and $Q_1, \ldots, Q_N$.

In an ideal setting, the structural modeler would like to use this information to estimate firm-level demand and cost specifications, such as those discussed in Sections 5–8. Unlike these previous models, however, assumptions about firms' fixed costs will now play an important role in these models, as fixed costs help determine which set of firms will produce. Additionally, assumptions about the timing of firms' decisions and the amount of information they possess become critical. These assumptions are important because, unlike in previous models, they have a critical impact on whether the empirical model has a pure-strategy equilibrium and whether any pure-strategy equilibrium is unique. In what follows, we use a series of models advanced by Bresnahan and Reiss (1991a, 1991b) to highlight some of these issues and the strengths and weaknesses of structural models.[34]

Bresnahan and Reiss develop econometric models to explain the number of sellers in several different localized product markets (such as dental services, new car dealers and movie theaters). For each product, they model how the number of sellers in a town varies with the town's population, and other demand and cost variables. The goal of their work is to understand how technological, demand and strategic factors affect market structure and competition. Like the airline example, they propose to do this by estimating how much demand it takes to support different numbers of firms. Unlike the airline example, however, the authors only have information on the number of firms in each market and their identities $a = (a_1, \ldots, a_M)$; they do not have price or quantity information. Thus, absent a structural model, the best they can do is summarize the conditional joint distribution of entry decisions given industry and firm characteristics. Such an approach is not that dissimilar from that taken in Dunne, Roberts and Samuelson (1988). When developing a structural model, Bresnahan and Reiss must take into account the fact that entry and exit are discrete events. Thus, their structural models will not typically involve marginal conditions, such as those used in the models of Sections 5, 6 and 7. Instead, they must rely on threshold conditions for entrants' unobserved profits.

The threshold conditions that Bresnahan and Reiss use come from simple static, perfect-information entry games. An example of such a game is the standard two-firm, simultaneous-move entry game. The payoffs to the players in this game are:

	Stay out ($a_2 = 0$)		Enter ($a_2 = 1$)	
Stay out ($a_1 = 0$)	$\Pi_1(0,0)$	$\Pi_2(0,0)$	$\Pi_1(0,1)$	$\Pi_2(0,1)$
Enter ($a_1 = 1$)	$\Pi_1(1,0)$	$\Pi_2(1,0)$	$\Pi_1(1,1)$	$\Pi_2(1,1)$

where the $\Pi_k(a_1, a_2)$ represent the profits firm k earns when firm 1 plays a_1 and firm 2 plays a_2 (a zero denotes the action "Stay Out" and a one denotes "Enter"). In most

[34] See also the work of Berry (1992) and other references cited in Berry and Reiss (in press).

textbook examples, the numbers in the payoff matrix are hypothetical. The economist then adds assumptions about players' information and a solution concept.

Bresnahan and Reiss' structural models build on this strategic representation of an entry game. Their econometric models postulate that the researcher observes the players' equilibrium action(s) in each sample market (e.g., $a_1 = 0$ and $a_2 = 1$) but does not observe the firms' economic profits (the $\Pi_k(0, 1)$). The logic of their models is to use a specific equilibrium solution concept to work backward from the observed equilibrium action(s) to statements about unobserved profits. Thus, the "structure" in their structural model are the economic and stochastic assumptions that allow them to go from discrete data to statements about continuous-valued profits. It should not be too surprising given our discussions in Sections 5–9, that Bresnahan and Reiss will have to introduce considerable structure in order to draw inferences about firm profits and behavior from discrete outcomes.

10.3. Modeling profits and competition

To understand the process by which Bresnahan and Reiss work from firms' observed actions back to statements about firms' unobserved profits, and to see what one can hope to estimate, it is useful to work with a specific entry model. To keep matters simple, imagine that we are modeling the number of symmetric firms, N, that produce a homogeneous good. The goal of the empirical analysis is to use the information in the zero-one entry indicators $a_1, a_2, \ldots, a_M$ of the $M \geqslant N$ potential entrants to draw inferences about firms' profit functions, i.e.,

$$\Pi_k(a_1, a_2, \ldots, a_M, X, Z, W, \theta). \tag{147}$$

Here X, Z, and W represents exogenous observables affecting demand and costs, and θ represents parameters of the profit function (e.g., demand and cost function parameters) that we wish to estimate. While the firms' profit functions could in principle include prices and quantities, Bresnahan and Reiss do not have this information. They thus are forced to work with profit functions where these endogenous variables have been substituted out.

The first step in the modeling process is to use assumptions about demand, costs and how firms compete to derive the functional form of Equation (147). Here Bresnahan and Reiss are helped by the presumption that if a potential entrant does not enter, it likely will earn zero profit – regardless of what the other potential entrants do. If firm i does enter, its profits depend on the number of other firms that enter (as summarized in the a_j). The exact way in which the number of other firms affects profits depends on what one assumes about demand, costs and competition. If, for example, firms have the same constant marginal cost c, have fixed costs of F, compete as Cournot competitors, and market demand is $p = \alpha - bQ$, then one can show

$$\Pi_k(a_1, a_2, \ldots, a_M, Z, \theta) = b\left(\frac{S}{\sum_{j=1}^{M} a_j + 1}\right)^2 - F, \tag{148}$$

where $S = (\alpha - c)/b$ is a measure of the potential size of the market. For firm i to have entered along with $N - 1$ other firms it must be the case that $\Pi_i \geqslant 0$. Similarly, if there is free entry, then it must be that the $(N + 1)$st entrant found it unprofitable to enter. These two bounds imply

$$\frac{S^2}{(N+1)^2} \geqslant \frac{F}{b} \geqslant \frac{S^2}{(N+2)^2}.$$

These inequalities provide useful information. For instance, if we know or could estimate the size of the market S and the slope of demand b, then we can place a bound on firms' unobserved fixed costs. While it is plausible to imagine having external measures of the market's size, S, it is much less likely one would have prior information about b. One solution would be to use price and quantity data to estimate b, yet this is exactly the problem that Bresnahan and Reiss have – they do not have price and quantity information.

The question then is what can one infer about demand and cost conditions from a cross section of markets? Bresnahan and Reiss' idea is to use information on the number of firms in very small to very large markets to estimate a sequence of so-called entry thresholds. These thresholds are a simple transformation of the market sizes $S_1, S_2, \ldots$ above, where S_i represents the size of the market just needed to support i firms. While the entry threshold levels are of limited use, their ratios are revealing. For example, if we take the ratio of the duopoly to the monopoly entry threshold assuming firms are Cournot competitors we get

$$\frac{S_2^2}{S_1^2} = \frac{9}{4} = 2.25. \tag{149}$$

That is, we should observe a second firm entering at 2.25 the size of the market required to support one firm. Similar calculations can be done for entry threshold ratios involving higher numbers of identical firms.

Of course, we need not observe the estimated (or observed) duopoly-monopoly threshold ratio equal to 2.25 (or the higher-order ratios consistent with this symmetric Cournot model). The question then is what should we infer? The answer is that economic theory can provide some suggestions. We can consider, for example, what happens when we change the assumption about how the duopolists compete. If the second entrant expects the monopolist to collude with it after entry, then the duopoly to monopoly ratio would equal 2.0. The three-firm to monopoly entry threshold ratio would be 3.0, and so on. Alternatively, if the second firm expected perfect competition (or Bertrand competition) post entry, we would never observe the second firm enter this natural monopoly. Thus, we can see that the degree of competition affects the entry threshold ratio. While we might be tempted to think the entry threshold ratio then is indicative of the degree of competition, with larger ratios suggesting more competition post entry, this is only true if we maintain our other assumptions. If, for example, we had used a quadratic cost function with increasing marginal costs, we also would see

changes in the entry threshold ratios as minimum efficient scale changes [see Bresnahan and Reiss (1991a)].

This last point brings us back to a point we made in the introduction: inferences in structural models typically depend heavily on maintained functional form assumptions. We often do not have the data to test these assumptions. In this application, for example, the absence of price and quantity data considerably limit what we can infer. Does this suggest that this structural model has little value because we have to make untestable assumptions? Our answer is no. The model has value because it makes clear what one can and cannot infer from the data. It also points future research toward what it is that one needs to observe to draw sharper inferences.

10.4. The econometric model

Our discussion so far has largely been based on an economic model with symmetric firms. We have yet to introduce stochastic assumptions or discuss the more realistic cases where there are observed and unobserved differences among firms. These additions introduce further complexities.

Recall that the data Bresnahan and Reiss have are the number of potential entrants M, the number (and possibly the identities) of the actual entrants, and demand and cost variables. Starting from primitive demand and cost function assumptions, they build a model of firms' equilibrium profits, which consists of a variable profit and a fixed cost term

$$\bar{\Pi}_k(a, Z, \theta) = \mathrm{VP}_i(a, Z, \theta) - F_i(a, Z, \theta). \tag{150}$$

Here, a is a vector describing the M potential entrants' entry actions, VP denotes variable profits, F fixed costs and i subscripts potential entrants. Although this expression depends on observable variables, the econometrician does not typically observe everything the firm does. Following the discrete choice literature popularized by McFadden, Heckman, and others, we might simply add an error term, ϵ, to profits to account for what we do not observe. Notice, however, that by assuming that the error is additive, we have placed structure on what it is about profits that the econometrician does not observe. Specifically, whatever it is that the econometrician does not observe, it enters the firms' optimal choices of prices and quantities in such a way that we obtain an additive error in Equation (150). What types of unobservables do and do not fit this specification? If we assume that the firms have unobserved differences in their constant marginal costs, then we will not obtain an additive error specification. On the other hand, if we assume that firms have different fixed costs, then we will. (This is because the marginal conditions for prices or quantities do not depend on the unobservable fixed cost.) Thus, while it is possible to justify the unrestricted additive structure in (150), it may make more economic sense to entertain alternative stochastic specifications for profits.

Assuming that the unobserved portion of profits is additive, we are now in a position to write down expressions for the equilibrium threshold conditions on firm profits. Following the discrete choice literature, we might consider modeling entry as the event that

the firm i's latent profits exceeds 0, or

$$\mathrm{VP}_i(a, Z, \theta) - \tilde{F}_i(a, Z, \theta) \geqslant \epsilon_i(a), \tag{151}$$

where the tilde above fixed costs denotes fixed cost up to an additive mean zero error. This model looks like a standard threshold condition in a conventional discrete choice model. The key difference is that the threshold conditions in the entry model contain the endogenous a_i variables. In other words, unlike in the standard discrete choice model, here agents' discrete decisions are interrelated. We therefore have to model simultaneously the N potential entrants' threshold conditions. This is the source of additional complications.

There is some precedent in the discrete choice literature for threshold conditions that include dummy endogenous variables (the a_i). For example, the household labor supply literature sometimes descriptively models the dependence of a household head's labor supply decision on their spouse's labor supply decision. Amemiya (1974) and others have studied the econometric properties of latent variable models that include dummy endogenous variables. Heckman (1978) introduced a systematic formulation of linear dummy endogenous variable models and discussed a variety of econometric issues associated with the formulation and estimation of such models. In particular, he and others have noted that arbitrary specifications of dummy endogenous variable models can lead to "coherency" and identification problems.

Bresnahan and Reiss showed that one could use the economic structure of discrete games to produce structural choice models with Heckman's econometric structure. Moreover, the identification issues that arise in Heckman's models often have natural economic interpretations. To see some of the connections, let us return to the normal form entry game above. Recall that the idea of Bresnahan and Reiss is to draw inferences about the unobserved payoffs from the observed equilibrium actions of the entrants. To link the observed actions to the payoffs, we employ an equilibrium solution concept. An obvious one to employ in analyzing an entry game is that of a Nash equilibrium. An outcome $\{a_1^*, a_2^*\}$ of the entry game is a Nash equilibrium if

$$\Pi_1\big(a_1^*, a_2^*\big) \geqslant \Pi_1\big(a_1, a_2^*\big) \quad \text{and} \quad \Pi_2\big(a_1^*, a_2^*\big) \geqslant \Pi_2\big(a_1^*, a_2\big) \tag{152}$$

for any a_1 and a_2. To make clear the connection between the Nash equilibrium outcomes and payoffs, we can rewrite the two-by-two entry game as:

	Stay out ($a_2 = 0$)	Enter ($a_2 = 1$)
Stay out ($a_1 = 0$)	$\Pi_1(0,0) \quad \Pi_2(0,0)$	$\Pi_1(0,1) \quad \Pi_2(0,0) + \Delta_0^2$
Enter ($a_1 = 1$)	$\Pi_1(0,0) + \Delta_0^1 \quad \Pi_2(1,0)$	$\Pi_1(0,1) + \Delta_0^1 + \Delta_1^1 \quad \Pi_2(1,0) + \Delta_0^2 + \Delta_1^2$

where the Δ's represent the incremental profits to each firm of entry. From the definition of a Nash equilibrium and the above payoff matrix we can deduce

$$a_1 = 0 \quad \Longleftrightarrow \quad \Delta_0^1 + a_2 \Delta_1^1 \leqslant 0,$$

$$a_2 = 0 \quad \Longleftrightarrow \quad \Delta_0^2 + a_1 \Delta_1^2 \leqslant 0. \tag{153}$$

These conditions link the observed actions to profits. Specifically, they tell us that all that the econometrician can infer from the observed equilibrium actions are statements about the Δ terms. In the case of a Nash equilibrium, we see this means that the econometrician cannot estimate $\Pi_1(0, 1)$ and $\Pi_2(1, 0)$, which are the profits the firms earn when it is out of the market. This makes perfect sense, as we can only learn about profits when a firm enters. To understand what we can estimate, it is useful to analyze the Δ's. The Δ_0^i term are the incremental profits that firm i earns in a monopoly. We might naturally think of this incremental profit as monopoly variable profits minus fixed costs, net of opportunity costs. The Δ_1^i terms are the profits that firm i gains (loses) relative to its incremental monopoly profit when it enters its competitor's monopoly market. This profit is most naturally thought of as the loss in variable profit from moving from a monopoly to a duopoly.

From assumptions about the structure of demand and costs, we can relate the incremental profit terms to underlying demand and cost variables and parameters. For example, in the symmetric linear demand and cost Cournot example, where $\Pi_i(0, 0) = 0$ we have

$$\begin{aligned} \Delta_0^i &= \frac{(\alpha - c)^2}{4b} - F = g(\alpha, c) - F, \\ \Delta_1^i &= \frac{5(\alpha - c)^2}{36b} = h(\alpha, c). \end{aligned} \tag{154}$$

Knowing this relationship between the Δ's and the underlying economic parameters, we can proceed to add error terms to the model to generate stochastic specifications. Assuming $F_i = F + \epsilon_i$ gives the following latent variable system

$$a_i = \begin{cases} 1 & \text{if } y_i^* = g(\alpha, c) - F + a_j h(\alpha, c) - \epsilon_i \geqslant 0, \\ 0 & \text{if } y_i^* = g(\alpha, c) - F + a_j h(\alpha, c) - \epsilon_i < 0, \end{cases} \tag{155}$$

for $i = 1, 2$ and $i \neq j$. This system bears a resemblance to Heckman's (1978) linear dummy endogenous variable systems. For instance, if we ignore the demand and cost parameters in $g(\cdot)$ and $h(\cdot)$, assume Δ_1^i is a constant, and $\Delta_0^i = X\beta_i$, where X is a vector of observable variables and β_i is a vector of parameters, then we obtain the linear dummy endogenous variable system

$$a_i = \begin{cases} 1 & \text{if } y_i^* = X\beta_i + a_j \delta - \epsilon_i \geqslant 0, \\ 0 & \text{if } y_i^* = X\beta_i + a_j \delta - \epsilon_i < 0. \end{cases} \tag{156}$$

Amemiya, Heckman, Maddala and others have noted we cannot estimate the above systems in general if the errors have unbounded support. The reason for this is that the reduced form is not always well defined for all values of the errors. Bresnahan and Reiss show that this econometric problem has a natural economic interpretation: namely, it is indicative of two types of problems with the underlying game. First, if the errors are unrestricted. the underlying game may have multiple pure-strategy equilibria.

Second, the underlying game may have no pure-strategy equilibria. These existence and uniqueness problems cause havoc with pure-strategy reduced forms.

One proposed solution to these problems is to assume that the model is recursive. This econometric solution, however, has unattractive economic implications for an entry game. Specifically, it amounts to assuming that a competitor's entry into a monopoly market does not affect the monopolist's profits. Thus, while this assumption is computationally attractive, it is economically and empirically unrealistic.

Bresnahan and Reiss go on to suggest how one can impose restrictions on profits that remove existence problems. They also suggest a solution for the nonuniqueness problem, which is to aggregate the nonunique outcomes (in this case the nonunique outcomes occur when one firm or the other firm could be a profitable monopolist) to obtain an economic model *of the number of firms in the market*, rather than a model *of which firms are in the market*. Bresnahan and Reiss also explore how changing the solution concept for the entry model changes the econometric structure of the game. The main one they explore is how changing the game from simultaneous-move Nash to sequential-move Stackleberg. In the latter case, the entry game generically has a unique equilibrium. The econometric model of this equilibrium also has a threshold interpretation, but it is more complicated than the simple linear structure above.

10.5. Estimation

Turning now to estimation, Bresnahan and Reiss (1991a) propose maximum likelihood methods for estimating the parameters of profits. In their empirical work, they focus on estimating models where the number of potential entrants is small. A key assumption in their work is that they actually know the number of potential entrants, and therefore the number of threshold conditions to impose. In much of their work, they ignore systematic differences in firms' profits and focus instead on modeling the number of firms that will enter geographically distinct markets. In particular, Bresnahan and Reiss assume that the demand for the products they look at is proportional to a town's current and future population size, and that the per capita demands for these products does not depend on population. This allows them to express market demand as $Q = D(Z, P)S$, where S is the "size" of the market. To simplify the analysis, Bresnahan and Reiss assume that sellers are the same, apart from potential differences in fixed costs.

Using these assumptions, Bresnahan and Reiss derive expressions for equilibrium monopoly and duopoly profits as a function of the size of the market S, other demand variables and cost variables. A key observation is that the size of the market S enters linearly into firm profits. Assuming there are only two possible entrants, firm 1 has post-entry profits

$$\Pi_i(1, a_2) = \big(g(Z, \beta) + a_2 h(Z, \delta)\big)S - F(a_2) - \epsilon. \tag{157}$$

From this relation, Bresnahan and Reiss identify entry thresholds for a monopolist and a duopoly. That is, the entry thresholds equal

$$S(a_2) = \frac{F(a_2) - \epsilon}{g(Z, \beta) + a_2 h(Z, \delta)}. \tag{158}$$

The entry thresholds are of interest because they tell us something about unobserved fixed costs relative to the variable profit parameters. While in principle, Bresnahan and Reiss should motivate the functions $h(Z, \delta)$ and $g(Z, \beta)$ from a specific model of demand and variable costs, in their empirical work they assume that these functions are linear in the Z variables (or constants). Bresnahan and Reiss make these assumptions both to simplify estimation and because they cannot easily separate cost and demand variables.

In most of their work, Bresnahan and Reiss focus on estimating ratios of entry thresholds. In their model, the ratio of the monopoly to the duopoly entry threshold equals:

$$\frac{S(1)}{S(0)} = \frac{F(1)}{F(0)} \frac{g(Z, \beta)}{g(Z, \beta) + h(Z, \delta)}. \tag{159}$$

This expression shows that the ratio depends on the extent to which the second entrant has higher fixed costs than if it were a monopolist and the extent to which duopoly profits are less than monopoly profits (here $h(Z, \delta) < 0$). Bresnahan and Reiss estimate the left-hand side by first estimating the parameters of the profit functions (150) and then forming the ratio (159). They then draw inferences about competition based on maintained demand and cost assumptions, much as we have discussed above. For example, they observe that entry threshold ratios in several different product markets are not dramatically different from that implied by a model where firms act as Cournot competitors. Again, however, their inferences about product market competition rest heavily on their assumptions about demand and costs, and they only explore a limited set of alternative demand and cost assumptions.

10.6. Epilogue

In Section 4 we stated that a structural modeling exercise should not go forward without a clear justification, in terms of economically meaningful magnitudes that can be estimated, for the many untestable assumptions necessary to specify and estimate a structural model. In this case, the justification for the structural model is its ability to recover estimates of the entry thresholds and the fixed costs of entry from the number of firms in a market. Neither of these magnitudes are directly observable and thus can be inferred after the researcher has made assumptions about the form of demand and firm-level costs, including entry costs. In contrast to the literature described in Sections 5 through 7 that uses market prices and quantities, with fewer observable market outcomes, these models rely more heavily on functional form and distributional assumptions to recover magnitudes of economic interest.

A number of researchers have extended Bresnahan and Reiss' models and explored alternatives [see Berry and Reiss (in press)]. In many respects these models share a common feature: to draw economic inferences from qualitative data on entry and exit, they have to impose considerable economic structure and in many cases sacrifice realism to obtain empirically tractable specifications. So what does this say about IO economists' progress in developing structural models of oligopolistic market structure? The bad news is that the underlying economics can make the empirical models extremely complex. The good news is that the attempts so far have begun to define the issues that need to be addressed. They also have clarified why simple probit models and the like are inadequate for modeling entry and exit decisions.

11. Ending remarks

More than fifty years ago, members of the Cowles Commission began a push to estimate empirical models that combined economic models with probability models. They labeled this enterprise econometrics. In the intervening years, some economists have come to think of econometrics as high-tech statistics applied to economic data. That is, that econometrics is a field that mainly focuses on the development of statistical techniques. While this may be true of some of econometrics, much of the Cowles Commission's original vision is alive and well. In this chapter, we have tried to provide a sense of how structural modeling proceeds in industrial organization. We used "structural econometric modeling" as opposed to "econometric modeling" in our title to emphasize that an application's setting and economics should motivate specific probability models and estimation strategies, and not the other way around.

We began by comparing nonstructural or descriptive, and structural models. We should emphasize once more that we see great value in both descriptive and structural models. IO economists, for example, have learned much about the sources of competition from case studies of competition in specific industries. Our introductory sections tried to provide a sense of the benefits and costs associated with developing and estimating descriptive and structural models. An important benefit of a structural model is that it allows the researcher to make clear how economics affects the conditional distribution of the data. For example, we can always regress market quantity on price, but this does not necessarily mean we have estimated the parameters of a market demand function. To know whether we have or have not, we need to be clear about supply and the sources of error in the estimating equation.

While economic theory can help guide the specification and estimation of economic quantities, there is no simple recipe for developing structural econometric models. There are a variety of factors that make structural modeling difficult. First, economic theories often are sufficiently complex that it is difficult to translate them into estimable relations. In this case, structural modelers who opt to estimate simpler models often are subject to the criticism that their models are too naive to inform the theory. Second, structural modelers often lack data on all of the constructs or quantities in an economic theory.

The absence of relevant data can considerably complicate estimation and limit what it is that the researcher can estimate with the available data. Third, economic theory rarely delivers all that the structural modeler needs to estimate a model. Much is left to the modeler's discretion. The structural modeler typically must pick: functional forms; decide how to measure theoretical constructs; decide whether to include and how to include variables not explicitly part of the theory; how to introduce errors into the model; and decide on the properties of errors. Each of these decisions involve judgments that cannot be tested. Thus, these maintained assumptions need to be kept in mind when interpreting structural model estimates, parameter tests and performing counterfactual calculations.

In our selective tour, we have tried to provide a sense of how IO researchers have dealt with some of these issues. Our intent was not to be a comprehensive review of all that has been done on a particular topic, but rather to provide a vision for some of the general modeling issues IO researchers face in linking IO theories to data. We hope that our chapter has conveyed a sense of progress, and also a sense that much remains for IO economists to explore.

References

Ackerberg, D., Rysman, M. (2005). "Unobserved product differentiation in discrete choice models: Estimating price elasticities and welfare effects". RAND Journal of Economics 36 (4), 771–788.

Amemiya, T. (1974). "Multivariate regression and simultaneous equation models when the dependent variables are truncated normal". Econometrica 42 (6), 999–1012.

Andersen, E.B. (1970). "Asymptotic properties of conditional maximum likelihood estimation". Journal of the Royal Statistical Society, Series B 32 (2), 283–301.

Applebaum, E. (1982). "The estimation of the degree of oligopoly power". Journal of Econometrics 19, 287–299.

Athey, S., Haile, P.A. (2002). "Identification of standard auction models". Econometrica 70 (6), 2107–2140.

Bajari, P., Benkard, L. (2001). "Discrete choice models as structural models of demand: Some economic implications of common approaches". Working manuscript. Stanford Graduate School of Business.

Bajari, P., Benkard, L. (2005). "Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach". Journal of Political Economy 113 (6), 1239–1276.

Baker, J.B., Bresnahan, T.F. (1988). "Estimating the residual demand curve facing a single firm". International Journal of Industrial Organization 6 (3), 283–300.

Baron, D.P. (1989). "Design of regulatory mechanisms and institutions". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam.

Baron, D.P., Besanko, D. (1984). "Regulation, asymmetric information and auditing". RAND Journal of Economics 15 (4), 447–470.

Baron, D.P., Besanko, D. (1987). "Monitoring, moral hazard, asymmetric information, and risk-sharing in procurement contracting". Rand Journal of Economics 18 (4), 509–532.

Baron, D.P., Myerson, R. (1982). "Regulating a monopolist with unknown costs". Econometrica 50 (4), 911–930.

Becker, G.S. (1962). "Irrational behavior and economic theory". Journal of Political Economy 70 (1), 1–13.

Berry, S.T. (1992). "Estimation of a model of entry in the airline industry". Econometrica 60 (4), 889–917.

Berry, S.T. (1994). "Estimating discrete-choice models of product differentiation". RAND Journal of Economics 25 (2), 242–262.

Berry, S.T. (2001). "Estimating the pure hedonic choice model". Working manuscript. Yale Department of Economics.

Berry, S.T., Levinsohn, J., Pakes, A. (1995). "Automobile prices in market equilibrium". Econometrica 63 (4), 841–890.

Berry, S.T., Levinsohn, J., Pakes, A. (2004). "Estimating differentiated product demand systems from a combination of micro and macro data: The new car model". Journal of Political Economy 112 (1), 68–105.

Berry, S.T., Linton, O., Pakes, A. (2004). "Limit theorems for estimating the parameters of differentiated product demand systems". Review of Economic Studies 71, 613–654.

Berry, S.T., Reiss, P.C. (2003). "Empirical models of entry and exit". In: Porter, R.H., Armstrong, M. (Eds.), Handbook of Industrial Organization, vol. 3. North-Holland, Amsterdam. In press.

Besanko, D. (1984). "On the use of revenue requirements regulation under imperfect information". In: Crew, M.A. (Ed.), Analyzing the Impact of Regulatory Change in Public Utilities. Lexington Books, Lexington.

Blackorby, C., Primont, D., Russell, R.R. (1978). Duality, Separability and Functional Structure. North-Holland, Amsterdam.

Borenstein, S. (1992). "The evolution of US airline competition". Journal of Economic Perspectives 6 (2), 45–73.

Bresnahan, T.F. (1981). "Departures from marginal-cost pricing in the American automobile industry: Estimates for 1977–1978". Journal of Econometrics 11, 201–227.

Bresnahan, T.F. (1982). "The oligopoly solution concept is identified". Economics Letters 10 (1–2), 87–92.

Bresnahan, T.F. (1987). "Competition and collusion in the American automobile market: The 1955 price war". Journal of Industrial Economics 35 (4, June), 457–482.

Bresnahan, T.F. (1989). "Empirical methods for industries with market power". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam.

Bresnahan, T.F. (1997). "Comment". In: Bresnahan, T.F., Gordon, R. (Eds.), The Economics of New Goods. University of Chicago Press, Chicago.

Bresnahan, T.F., Reiss, P.C. (1985). "Dealer and manufacturer margins". RAND Journal of Economics 16 (2), 253–268.

Bresnahan, T.F., Reiss, P.C. (1991a). "Entry and competition in concentrated markets". Journal of Political Economy 99 (5), 977–1009.

Bresnahan, T.F., Reiss, P.C. (1991b). "Empirical models of discrete games". Journal of Econometrics 48 (1–2), 57–81.

Brueckner, J.K., Dryer, N.J., Spiller, P.T. (1992). "Fare determination in hub and spoke networks". RAND Journal of Economics 23 (3), 309–323.

Camerer, C. (1995). "Individual decision making". In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton University Press, Princeton.

Campo, S., Guerre, E., Perrigne, I.M., Vuong, Q. (2003). "Semiparametric estimation of first-price auctions with risk-averse bidders". Working manuscript. University of Southern California.

Christensen, L.R., Greene, W.H. (1976). "Economics of scale in US electric power generation". Journal of Political Economy 84 (4), 655–676.

Corts, K.S. (1999). "Conduct parameters and the measurement of market power". Journal of Econometrics 88 (2), 227–250.

Dalen, D.M., Gomez-Lobo, A. (1997). "Estimating cost functions in regulated industries characterized by asymmetric information". European Economic Review 41 (3–5), 935–942.

Davis P. (2000). "Demand models for market-level data". Working manuscript. MIT Sloan School.

Deaton, A., Muellbauer, J. (1980). Economic and Consumer Behavior. Cambridge University Press, Cambridge.

Dixit, A.K., Stiglitz, J.E. (1977). "Monopolistic competition and optimum product diversity". American Economic Review 67 (3), 297–308.

Dunne, T., Roberts, M.J., Samuelson, L. (1988). "Patterns of firm entry and exit in US manufacturing industries". RAND Journal of Economics 19 (4), 495–515.

Engel, E. (1857). "Die Productions-und Consumptionsverhältnisse des Königreichs Sächsen". In: Zeitschrift des Statischen Bureaus des Königlich Söchsischen Ministeriums des Inneren, Nos. 8 and 9.

Evans, D., Heckman, J.J. (1984). "A test of subadditivity of the cost function with an application to the Bell system". American Economic Review 74 (4), 615–623.
Gagnepain, P., Ivaldi, M. (2002). "Incentive regulatory policies: The case of public transit systems in France". RAND Journal of Economics 33 (4), 605–629.
Goldberg, P.K. (1995). "Product differentiation and oligopoly in international markets: The case of the US automobile industry". Econometrica 63 (4), 891–951.
Goldberger, A.S. (1991). A Course in Econometrics. Harvard University Press, Cambridge.
Gollop, F.M., Roberts, M.J. (1979). "Firm interdependence in oligopolistic markets". Journal of Econometrics 10 (3), 313–331.
Gorman, W.M. (1959). "Separable Utility and Aggregation". Econometrica 27 (3), 469–481.
Gorman, W.M. (1970). "Two-stage budgeting". In: Blackorby, C., Shorrocks, A. (Eds.), Separability and Aggregation. In: Collected Works of W.M. Gorman, vol. 1. Clarendon Press, Oxford.
Green, E.J., Porter, R.H. (1984). "Noncooperative collusion under imperfect price information". Econometrica 52 (1), 87–100.
Guerre, E., Perrigne, I.M., Vuong, Q. (2000). "Optimal nonparametric estimation of first-price auctions". Econometrica 68 (3), 525–574.
Haavelmo, T. (1944). "The probability approach in economics". Econometrica 12 (Suppl.), iii-vi and 1-115.
Haile, P.A., Tamer, E. (2003). "Inference with an incomplete model of English Auctions". Journal of Political Economy 111 (1), 1–51.
Hanemann, W.M. (1984). "Discrete/continuous models of consumer demand". Econometrica 52 (3), 541–562.
Härdle, W. (1990). Applied Nonparametric Regression. Cambridge University Press, London.
Härdle, W., Linton, O. (1994). "Applied nonparametric methods". In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.
Hausman, J. (1997). "Valuation of new goods under perfect and imperfect competition". In: Bresnahan, T.F., Gordon, R. (Eds.), The Economics of New Goods. University of Chicago Press, Chicago.
Hausman, J., Leonard, G., Zona, D. (1994). "Competitive analysis with differentiated products". Annales d'Economie et de Statistique 34 (0), 159–180.
Heckman, J.J. (1978). "Dummy endogenous variables in a simultaneous equation system". Econometrica 46 (4), 931–959.
Hendricks, K., Paarsch, H.J. (1995). "A survey of recent empirical work concerning auctions". Canadian Journal of Economics 28 (2), 403–426.
Hendricks, K., Porter, R.H. (1988). "An empirical study of an auction with asymmetric information". American Economic Review 78 (5), 865–883.
Hendricks, K., Porter, R.H. (2000). "Lectures on auctions: An empirical perspective". In: Armstrong, M., Porter, R.H. (Eds.), Handbook of Industrial Organization, vol. 3. North-Holland, Amsterdam. In press.
Hood, W.C., Koopmans, T.C. (1953). Studies in Econometric Method, Cowles Commission Monograph no. 14. John Wiley, New York.
Krasnokutskaya, E. (2002). "Identification and estimation of auction models under unobserved auction heterogeneity". Working manuscript. Yale Department of Economics.
Laffont, J.J. (1997). "Game theory and empirical economics: The case of auction data". European Economic Review 41 (1), 1–35.
Laffont, J.J., Tirole, J. (1986). "Using cost observation to regulate firms". Journal of Political Economy 94 (3), 614–641.
Laffont, J.J., Tirole, J. (1993). A Theory of Incentives in Procurement and Regulation. MIT Press, Cambridge.
Laffont, J.J., Vuong, Q. (1996). "Structural analysis of auction data". American Economic Review 36 (2), 414–420.
Laffont, J.J., Ossard, H., Vuong, Q. (1995). "Econometrics of first-price auctions". Econometrica 63 (4), 953–980.
Lau, L.J. (1982). "On identifying the degree of industry competitiveness from industry price and output data". Economics Letters 10 (1–2), 93–99.
Lerner, A. (1934). "The concept of monopoly and the measurement of monopoly power". Review of Economic Studies 1 (3), 157–175.

Li, T., Perrigne, I.M., Vuong, Q. (2002). "Structural estimation of the affiliated private value auction model". RAND Journal of Economics 33, 171–193.
Lindh, T. (1992). "The inconsistency of consistent conjectures". Journal of Economic Behavior and Organization 18 (1), 69–90.
Maddala, G.S. (1983). Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.
McAfee, R.P., McMillan, J. (1987). "Auctions and bidding". Journal of Economic Literature 25 (2), 699–738.
Milgrom, P.R., Weber, R.J. (1982). "A theory of auctions and competitive bidding". Econometrica 50 (5), 1089–1122.
Morrison, S.A., Winston, C. (1996). "The causes and consequences of airline fare wars". Brookings Papers on Economic Activity: Microeconomics, 85–123.
Nevo, A. (2000). "Mergers with differentiated products: The case of the ready-to-eat cereal industry". RAND Journal of Economics 31 (3), 395–421.
Ott, J. (1990). "Justice Dept. Investigates Carriers' Pricing Policies". Aviation Week and Space Technology 133 (3), 18–20.
Paarsch, H.J. (1992). "Deciding between the common and private values paradigms in empirical models of auctions". Journal of Econometrics 51 (1–2), 191–216.
Paarsch, H.J. (1997). "Deriving an estimate of the optimal reserve price: An application to British Columbia timber sales". Journal of Econometrics 78 (2), 333–357.
Petrin, A. (2002). "Quantifying the benefits of new products: The case of the minivan". Journal of Political Economy 110 (4), 705–729.
Phillips, A.W. (1958). "The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957". Economica 25 (100), 283–299.
Pinkse, J., Slade, M., Brett, C. (2002). "Spatial price competition: A semiparametric approach". Econometrica 70 (3), 1111–1155.
Pollak, R.A., Wales, T.J. (1992). Demand System Specification and Estimation. Oxford University Press, New York.
Porter, R.H. (1983). "A study of cartel stability: The Joint Executive Committee, 1880–1886". Bell Journal of Economics 14 (2), 301–314.
Quandt, R. (1988). The Econometrics of Disequilibrium. Basil Blackwell, Oxford.
Riordan, M.H. (1985). "Imperfect information and dynamic conjectural variations". RAND Journal of Economics 16 (1), 41–50.
Rosse, J.N. (1970). "Estimating cost function parameters without using cost data: Illustrated methodology". Econometrica 38 (2), 256–275.
Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
Sims, C.A. (1980). "Macroeconomics and reality". Econometrica 48 (1), 1–47.
Spiller, P.T., Favaro, E. (1984). "The effects of entry regulation on oligopolistic interaction: The Uruguayan Banking Sector". RAND Journal of Economics 15 (2), 244–254.
Stigler, G. (1964). "A theory of oligopoly". Journal of Political Economy 72 (1), 44–61.
Ulen, T.S. (1978). "Cartels and regulation". Unpublished PhD dissertation. Stanford University.
White, H. (1980). "Using least squares to approximate unknown regression functions". International Economic Review 21 (1), 149–170.
Windle, R. (1993). "Competition at 'Duopoly' Airline Hubs in the US". Transportation Journal 33 (2), 22–30.
Wolak, F.A. (1994). "An econometric analysis of the asymmetric information, regulator–utility interaction". Annales d'Economie et de Statistique 34 (0), 13–69.
Wolak, F.A. (2000). "An empirical analysis of the impact of hedge contracts on bidding behavior in a competitive market". International Economic Journal 14 (2), 1–40.
Wolak, F.A. (2003). "Identification and estimation of cost functions using observed bid data: An application to electricity". In: Detwatripont, M., Hansen, L.P., Turnovsky, S.J. (Eds.), Advances in Econometrics: Theory and Applications. In: Eighth World Congress, vol. 2. Cambridge University Press, Cambridge.