Structural Econometric Modeling: Rationales and Examples from Industrial Organization

by

Peter C. Reiss
Graduate School of Business
Stanford University
Stanford, CA 94305-5015
preiss@optimum.stanford.edu

Frank A. Wolak
Department of Economics
Stanford University
Stanford, CA 94305-6072
wolak@zia.stanford.edu

Incomplete Draft — June 13, 2002

NOTE

Prepared for the *Handbook of Econometrics*. This is a work in progress. We welcome comments on substance, clarity and style.

Contents

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), page xv). Many economists today, however, view econometrics as a field primarily concerned with statistical issues rather than economic questions. This has led some to draw a distinction between *econometric modeling* and *structural econometric modeling*, the latter phrase being used to emphasize the original Cowles Foundation mission for econometric research.

This chapter has three main goals. The first is to explain both what structural econometric modeling is and to describe the basic elements of a structural econometric model. While the process of deriving a coherent statistical models from an economic model might seem so obvious that it would be routine, nothing could be further from the truth. Structural econometric models simultaneously must: (1) be flexible statistical descriptions of data; (2) respect the details of economic institutions under consideration; and (3) be sensitive to the non-experimental nature of economic data. Moreover, just because an empirical researcher includes errors in an economic model does not guarantee that the resulting statistical model will be coherent or realistic. In this chapter we illustrate the subtleties and difficulties involved in combining economic and statistical models.

The process of building a structural model involves a series of interrelated steps. These steps require empiricists to trade off economic and statistical assumptions. Making these tradeoffs is not straightforward. While most econometrics textbooks do an excellent job teaching statistical methods, few discuss how to combine economic and statistical models. Fewer discuss why researchers must trade off economic and statistical assumptions to construct a structural econometric model. As a consequence, there is little consensus among economists on how to build and interpret structural econometric models. Thus, a second goal of this chapter is to propose a general framework for developing and evaluating structural econometric models. Although some elements of this framework originated with the founders of the Econometric Society, we add elements that are specific to a field of interest to us – industrial organization (IO).

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 IO researchers. These models are used to 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 the above topics.

Rather, we seek to illustrate how IO researchers have used economic structure and statistical assumptions to identify and estimate economic quantities. Our hope is that in doing so we can provide a sense of the benefits and limitations of structural econometric models generally.

This chapter is organized as follows. We begin with several examples of what we mean by a structural econometric model. We go on to illustrate the strength and weaknesses of structural econometric models through an extended series of examples. These examples provide a context for our structural modeling framework. Following a description of this framework, we use the framework to evaluate structural models from industrial organization. First, we consider models of competition in homogeneous product markets in which researchers estimate "conduct" or competitive "conjectural variation" parameters. We then discuss structural approaches to modeling competition in differentiated product markets, private or asymmetric information in auctions and principal-agent relations, and models of discrete strategic actions, such as entry and exit.

2 Descriptive and Structural Modeling in Econometrics

Empirical work in economics can be divided into two general categories: descriptive and structural. Descriptive work has a long and respected tradition in economics. Pure descriptive work can proceed without any reference to an economic model. For example, economists measure such things as the unemployment rate without relying on a particular model of the determinants of unemployment. Descriptive work also is concerned with uncovering time series or cross-section patterns in economic variables. The primary goal of most descriptive work in economics is to uncover trends, patterns or associations that might stimulate additional theoretical or empirical analyses. 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 logarithms of total household expenditures has shaped subsequent theoretical and empirical work on household consumption behavior (see Deaton and Muelbauer (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.

Many researchers believe there is an intermediate category of empirical research, most often referred to as "reduced form" analysis. This term is usually used to signal an

econometric model in which the researcher presumes there is a linear association between a dependent and one or more independent variables. Economics plays a role in these regression models to the extent that it suggests which variables are dependent and which are independent. The term "reduced form" also is meant to signal to others that the independent variables are presumed exogenous and that the regression coefficients capture by how much the dependent variable will change if the independent variable changes by one unit – *holding everything else constant.* One of the main goals of this chapter is to argue that this use of the term reduced form is inappropriate and not what members of the Cowles Commission intended. These regressions are in fact descriptive, and not an intermediate category of empirical model.

Descriptive work in general is concerned with estimating 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;

$Q_\alpha(y \mid x)$ the α conditional quantile of y given x; or

$BLP(y \mid x)$, the best linear predictor (BLP) of y given x.

Recently, statisticians and econometricians have devoted substantial energy to devising flexible methods for estimating joint densities. For example, statisticians have proposed kernel density techniques and other data smoothing methods for estimating $f(x, y)$. See Silverman (1986) and Hardle (1990) for introductions to these procedures. Although these non-parametric estimation techniques have the advantage of allowing the researcher to estimate the joint density of x and y while imposing minimal restrictions, these methods have a cost. Specifically, to yield much precision, smoothing techniques typically require enormous amounts of data. Moreover, the amount of data required to obtain precise estimates grows rapidly with the dimension of x and y. Silverman describes this as a "curse of dimensionality." He presents calculations that suggest that economists will need hundreds of thousands of observations to have much faith in these flexible techniques for estimating $f(x, y)$.[1]

Even in those rare circumstances when a researcher has sufficient data to estimate $f(x, y)$ flexibly, there are still compeling rationales for preferring to estimate a structural econometric model. A structural econometric model imposes the additional restriction that the population joint distribution of x and y, $f(x, y)$, reflects the behavior of economic actors. This implies that observed data can be used to recover estimates of parameters of underlying economic primitives that are invariant to the en-

[1]Silverman (1986, Table 4.2) shows that 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.

vironment that the economic agents operate in. Thus, like descriptive work, structural econometric modeling is about characterizing the properties of the joint distribution of economic data. Unlike descriptive work, structural models seek to estimate economic parameters or primitives from the joint distribution of the data. The essential components of a structural model then are the theoretical and statistical assumptions that allow a researcher to recover economic primitives from data. These assumptions minimally must be economically realistic and statistically sound. For the structure to be realistic, it must reasonably describe the economic and institutional environments generating the data. For it to be coherent, it must be possible to recover structural parameter estimates from all plausible data realizations.

To understand the process of building a structural model, consider a researcher who wishes to use household consumption, price and demographic data to estimate household demand for a particular good. The first step in building a structural model is to show that economic theory places restrictions on the joint distribution of household consumption, prices and income. The structural modeler might start by assuming the existence of a household-level utility function $U(x, z, \theta)$ that is a function of consumption x, taste parameters θ, and household characteristics z. The modeler might then use consumer theory to derive a mathematical relationship between household consumption, prices p and household characteristics: $x = h(p, z, \theta)$. Of course this theory will not perfectly explain households' consumption. The researcher therefore must either enrich the economic model or introduce error terms that represent variables outside the economic theory. These error terms might represent unobserved differences among agents, agent optimization errors, or errors introduced during data collection. For example, a structural modeler might assume that he does not observe all of the household characteristics that determine differences in preferences. He could model this incompleteness explicitly by introducing a vector of observed household characteristics, ϵ directly into the household utility functions: $U = U(y, x, \theta, \epsilon)$. By maximizing household utility subject to the household's budget constraint we now obtain demand functions that depend on these unobserved characteristics: $y = h(x, \theta, \epsilon)$.

To estimate the unknown utility ("structural") parameters θ, the structural modeler would then add assumptions about the joint population distribution for the unobserved tastes and the x's. For example, he would assume a specific joint distribution for x and ϵ. From this joint distribution, he could apply a change of variables to derive the joint distribution of the observed data $f(x, y)$ or other objects such as $f(y|x)$. The critical question at this point is: can he now find a method for estimating θ from the structure of $f(x, y)$? Ideally, the researcher must demonstrate that his econometric model is consistent with the observed joint density of x and y and that he can consistently estimate θ using the available data.

To summarize this process, structural econometric modeling involves a sequence of economic and statistical assumptions that gives rise to a joint density of x and y. Examples of economic assumptions are: What utility function should be used? What

is the budget constraint faced by the consumer? Examples of stochastic assumptions are: What types of errors should be introduced and where should they be introduced? Do these stochastic assumptions characterize the complete distribution, or might estimation be based on a statistical object that can be derived from the complete distribution? In what follows, we discuss these and other choices that structural modelers make. We loosely group these choices into three main groups: economic, statistical and tractability assumptions.

In closing this section, we would like to emphasize the fundamental difference between structural and descriptive econometric models. The ultimate goal of descriptive work is to estimate the joint density of x and y, although most economic researchers focus on estimating best linear predictor functions or the conditional density of y given x. Consequently, a descriptive researcher cannot make claims about causation or economic behavior, because these cannot be inferred from the distribution of data alone. On the other hand, a structural modeler can recover estimates of economic magnitudes and determine the extent of causation, but only because he is willing to make the economic and statistical assumptions necessary to infer these magnitudes from his econometric model for the joint density of x and y. This is a major strength of a structural econometric model – by making clear what economic assumptions are required to draw specific economic inferences from the data, the structural modeler makes it possible for others to assess the plausibility and sensitivity of the findings to these assumptions. One might think that a descriptive modeler can recover unambiguous estimates of economic magnitudes without making these economic and statistical assumptions. However, as we show below, this is not in general possible. In order set the stage for this discussion, we first need to describe in more detail where "structure" comes from in structural econometric models.

3 Putting the '*Econ*' Back into *Econ*ometrics

3.1 Sources of Structure

There are two sources of "structure" in structural models. First, economic theories deliver mathematical statements about the relationship between x and y. These mathematical statements often are deterministic, and as such do not speak directly to the distribution of noisy economic data. It is the applied researcher who adds the second source of structure, which are statistical sampling and other stochastic assumptions that specify how data on x and y were generated. This second source is necessary to transform deterministic models of economic behavior into stochastic econometric models. Thus, the "structure" in structural models typically comes from both economics and statistics.

Varying degrees of economic and stochastic structure can be imposed. Purists believe

that structural models must come from fully-specified stochastic economic models. Others believe that it is acceptable to add structure if that structure facilitates estimation or allows the researcher to recover some economically meaningful parameters. For example, 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 in fact reasonable or true. Put another way, the "structure" in a structural model is there because the researcher chose explicitly or implicitly to put it there. One of the advantages of structural econometric models is that researchers can examine the sensitivity of structural models and estimators to alternative economic and statistical assumptions. This is, however, easier said than done.

To illustrate how economists can introduce structure into an econometric model, we begin by examining two stylized econometric models. The purpose of the first model is to illustrate the difference between a descriptive and a structural model. This example shows that the same linear regression model can be a descriptive or a structural model depending on what economic and statistical assumptions the researcher is willing to make.

Example 1

We imagine an economist with a cross-section of firm-level data on output, Q_t, labor inputs, L_t, and capital inputs, K_t for firm t. To describe the relationship between output and inputs, the researcher might estimate the following linear regression by ordinary least squares (OLS):

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

where the θ's are unknown coefficients and the e_t is an error term that accounts for the fact that the right-hand side 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 $x_t = (1, ln(L_t), ln(K_t))'$. Goldberger (1991) Chapter 5 provides an excellent discussion of best linear predictors. The best linear predictor of y given a univariate x is $BLP(y \mid x) = a + bx$, where $a = E(y) - bE(x)$ and $b = Cov(y, x)/Var(x)$. Notice that the coefficients, a and b, of the best linear predictor function are statistical (and not economic) functions of the 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 positive definitive matrix, then we can show that 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 consistently estimate the linear function of the logarithm of labor and logarithm of capital that best predicts the logarithm of output.

Many economists, however, see regression (1) as being more than a descriptive regression. They would base their reasoning on the observation that (1) is essentially looks like a logarithmic restatement of a theoretical Cobb-Douglas production function: $Q_t = A\, L_t^{\alpha}\, K_t^{\beta}$. Because of the close resemblance, they would argue that (1) is in fact a "structural" and not a descriptive econometric model.

A critical missing step in this logic is that a Cobb-Douglas production function is deterministic relationship, 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 logarithm of the two inputs. In other words, it the combination of an economic assumption (production is truly Cobb-Douglas) and statistical assumptions (e_t satisfies certain moment conditions) that distinguishes a descriptive model from a structural model.

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 e_t was all measurement error? As we discuss later, this is likely too strong an assumption. A more plausible assumption is that the error, e_t, also includes unobservable (to the econometrician) differences in each firm's productivity (e.g., an unobservable component of A in the Cobb-Douglas function). This component of A is observed by the firm, before it makes its input choices. This economic assumption implies that e_t is correlated with observed input choices. This correlation necessitates using something other than OLS to recover consistent estimates of the parameters of the Cobb-Douglas production function.

Even if one were willing to assume that e_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.

This seemingly innocuous assumption of technologically efficient production may be inappropriate for some industries. There are a number of markets and environments where firms do not necessarily operate along their production function. 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. Failure to recognize all of the economic and statistical assumptions necessary to interpret estimated linear regression coefficients as parameters of an economic primitive is often why researchers claim to find regression results that are inconsistent with economic theory. ■

The following example illustrates the general process of constructing a structural econometric model by combining a deterministic economic theory with stochastic assumptions. As in the above example, a linear regression is estimated to recover the parameters of the underlying economic environment, but this statistical model is derived from a stochastic equilibrium economic model.

Example 2

Suppose an IO economist has firm output and cost data from different geographic monopoly markets where firms have constant, observable marginal costs c_i and face a linear inverse demand curve, $p = a - bq$, where a and b are constants and q is output. Static monopoly theory predicts that each monopolist's profit-maximizing output will be: $q_i = (a - c_i)/2b$. In other words, there will be a linear relation between each monopolist's output and its marginal cost, $q_i = \theta_0 + \theta_1 c_i$.

No IO economist would be so naive as to think that this static model will explain output data perfectly. Observed monopoly outputs will depart from this deterministic relation. This leaves the modeler with the choice of abandoning the theory or "fixing" it to rationalize the model's error. A common response is to presume that the theory and its associated functional forms are correct, but the empirical economist does not observe everything that the firm does. This amounts to the empirical modeler adding error terms into the deterministic model to account for differences between the theory and the data. Although we discuss this process more fully below, one way of doing this in the present model is to assume that the firms' demand intercepts differ in ways that are unobserved by the economist, and that these errors have the form $a_i = a + \epsilon_i$. This

specification leads to the stochastic (from the perspective of the researcher) relation between quantities and marginal costs $q_i = (a - c_i)/2b + \epsilon_i/2b = \theta_0 + \theta_1 c_i + \eta_i$. Note that this is not the only possible fix of this kind. This specification is indistinguishable from a model where the error comes from unobserved differences in firms' marginal costs. That is, assume $c_i^{True} = c_i^{Obs} + \epsilon_i$, where c_i^{Obs} are observed and c_i^{True} are true costs. Then $q_i = (a - c_i^{True})/2b = (a - c_i^{Obs})/2b + \epsilon_i/2b$.

While these error structures account for the deterministic economic model not perfectly predicting observed prices and quantities, they do not contain enough structure to justify a particular estimation method. This is where the empirical researcher must add assumptions. As we indicated in the first example, sometimes these assumptions are a matter of convenience. For example, a common assumption here is $E(\epsilon_i \mid c_i) = 0$, which also implies $E(\epsilon_i) = 0$. This assumption justifies using least squares to obtain consistent estimates of the conditional expectation function $E(q_i \mid c_i) = \theta_1 + \theta_2 c_i$ and consistent estimates of θ. From there, one can use the theory to recover estimates of the population average inverse demand intercept $a = E(a_i) = \theta_1/\theta_2$ and the slope of inverse demand $b = \theta_1/(2\theta_2)$. But what justifies the mean independence assumption for the error? The answer hopefully can be justified on economic or practical grounds; absent such answers, the assumption is one of statistical convenience. Is it bad to make statistical assumptions to facilitate estimation? Not necessarily, but the researcher must be clear that the assumption is critical to the consistency of the estimation results. ■

This example highlights the two steps necessary to construct a structural econometric model. First the researcher specifies an economic model of the phenomenon under consideration. Second she incorporates unobservables into the economic model. This second step of incorporating unobservables in economic models should receive significantly more attention than it does. This is both because the assumptions made about the unobservables will impact estimation and because not any old stochastic specification will do. For instance, for the stochastic specification to make sense, it trivially must be able to rationalize all possible realizations of the observed endogenous variables. Sections 4 and 5 illustrate the importance of stochastic specifications in more detail and illustrate potential pitfalls.

3.2 Why Use Structural Models?

We see three general reasons to go to the trouble of specifying and estimating a structural econometric model.

First, a structural model can be used to estimate unobservable economic parameters or behavioral responses from non-experimental data. In our second example, we used a profit-maximizing monopoly model to derive a relationship between output and

marginal cost. By combining this economic model with statistical assumptions about unobserved demand differences, we were able to find a way to recover consistent estimates of an average demand function. This function gives the amount consumers would purchase as a function of a market price, regardless of the number of producers in the market.

Other examples of behavioral or 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. If a researcher only observes market-clearing prices and quantities, and exogenous demand and supply shifters, the impact of a change in an exogenous variable on the unobserved amount demanded can only be recovered from a structural model. Conversely, a descriptive OLS regression can only tell us how the best linear predictor of an equilibrium price or quantity changes in response to a change in one of the regressors. No statements about the response of the quantity demanded or quantity supplied can be made using non-experimental data without the researcher specifying a structural econometric model of supply and demand.

Second, structural models can be used to simulate changes in equilibrium outcomes resulting from changes in the underlying economic environment. That is, one can use the estimated structure to predict what would happen if certain elements of the environment change. For instance, in our monopoly markets example, once we know the parameters of the demand curve and the firm-level marginal cost function, we can construct predictions about how the market equilibrium would change aspect of the economic environment changed. For example, we could predict by how much price would fall if an identical second firm entered the monopoly market. Economic theory predicts that if the two firms competed as Cournot duopolists, then each duopolist's output would be $q_i = (a - c_i)/3b$, which is two-thirds of the monopoly output. This is precisely the *structure* we need to obtain the joint density of firm-level output and marginal cost for the Cournot duopoly. We would first use consistent estimates of a and b to construct $f(q_i, c_i \mid \theta)$, the joint density of firm-level monopoly output and marginal cost. We would then compute the joint density of firm-level output and marginal under Cournot duopoly as: $f(q_i, c_i \mid \theta')$, where $\theta' = 2\theta/3$. It is economics that links the monopoly density parameters to those of a duopoly.

Another example of how we might use the monopoly model, is to ask what would happen if we regulated the monopolists' prices. Suppose in fact that the regulator sets the monopoly price equal to c_i. This would imply that $q_i = (a_i - c_i)/b$, which is twice the unregulated monopoly output. Consequently, $f(q_i, c_i \mid \theta')$ for the regulated monopoly case is equal to the unregulated monopoly density of q_i and c_i evaluated at the point $\theta' = 2\theta$.

These examples illustrate the benefits of structural modeling relative to descriptive modeling. To contrast the two approaches, notice that we could use flexible density

estimation techniques to estimate the joint density of $f(q_i, c_i)$ *in monopoly markets.* Because these flexible methods do not provide estimates of underlying economic parameters, they do not allow us to calculate how the density would change in markets for which we do not have data, such as duopoly or regulated markets. This underscores our point that unless a researcher is willing to make assumptions about the underlying economic environment, the only inferences that can be drawn from a descriptive analysis are those that pertain to the joint density of the data.

Third, 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, it is important to recognize that structural parameter estimates may well be very sensitive to these assumptions. For example, if we were trying to estimate consumer surplus from the demand estimates in Example 2, 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 because the "structure" in structural models forces researchers to grapple directly with the economic consequences of assumptions. As noted in the previous paragraph, if the applied researcher is unwilling to make any assumptions about the behavior of economic agents whose actions produce the observed data, only statements about the properties of best-linear predictor functions, conditional expectations, conditional quantiles and conditional densities are possible.

The "structure" in structural models also can affect the quality of 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 it usually is, economic theory is silent on which functional form and variable selection issues, researchers will be forced to make what can appear 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 hypothesis test. For instance, if a reseacher 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 impossible 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) there is a chance the effect of price may be exaggerated by the omission of relevant variables.

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 sensible changes in functional forms or the inclusion and exclusion of variables not closely tied to economic theory.

Finally, we should emphasize here that the advantages of structural models do not always favor structural models over descriptive 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 a substantial body of economic theory exists, empirical researchers should as much as possible shape assessments of that theory and its policy-making implications by estimating the foundations of the theory. These foundations include such primitives as consumers' utility functions and firms' production sets. Structural models have the added advantage of making it clear to theorists and other empiricists what assumptions must be made in order to link theory to data. By being clear about what economic theory and empirical models can and cannot address, it becomes easier for other researchers to evaluate and improve models.

These advantages of course do not 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 problems with how data are generated and collected (e.g., aggregation and censoring). Such complications may force the structural modeler to further simplify or limit models. 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 our first two examples.

3.3 Regressions and Structural Modeling

Empirical researchers often mistake statistical structure for economic structure. A leading example of this confusion is the way some economists interpret the results of a linear regression. For example, the model, $y = \alpha + x\beta + \epsilon$ lends an aura of "structure" as to how y is related to x. It suggests for example that x determines (or worse "drives") y. However, given a sample of T observations on these two variables, we can always regress y on x or x on y. The critical question is: What we do make of each set of coefficient estimates? Absent an economic model, the most we can say about the resulting coefficient estimates comes from statistics. As discussed earlier, so long as the first two sample moments of the joint distribution of x and y converge to their population analogues, each least squares regression yields consistent estimates of a best linear predictor function, in the former case the best linear predictor of y given x, $BLP(y|x)$, and in the latter case, $BLP(x|y)$.

In spite of this logic, many economists believe linear regression coefficients reveal economic "structure" even when there is no explicit economic model used to justify the estimated regression. For instance, in the 1960's and 1970's, IO economists were interested in whether firm profitability was related to the number and sizes of competitors in an industry. This led to many papers that estimated linear regressions of firm profitability (e.g., rates of return on capital) on market concentration measures (e.g., share of market output produced by the largest four firms) and other controls. Although these regressions regularly yielded positive estimated coefficients on market concentration, the coefficient magnitudes varied considerably from study to study. How did IO economists interpret these regression results? Many interpreted these regressions as supporting particular economic models of competition. Some argued that the positive coefficient reflected the ability of firms in highly concentrated markets to limit competition and thereby raise prices and profits. Others argued that, because more efficient firms have larger market shares, the results reflected the rents earned by efficient firms.[2]

This practice of using regression coefficient signs to corroborate economic models is widespread. It is supported by a belief that multiple regressions allow researchers to "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 regressing y on x and other variables. If the regression returns a positive coefficient on x, then this is seen by some as saying "the data support the economic theory." If x has the wrong sign, this sometimes leads the researcher to reject the theory; more often, it spurs the researcher to search for variables that could mediate the effect of x on y. Occasionally researchers experiment

[2]For an extended discussion of these models and the debate see for example Martin (1993), Chapters 17 and 18.

with alternative functional forms or different measures for y and x in their search for estimates that confirm a theory.

Critics of an economic theory play the same game. They suggest in seminars or referee reports that the positive effect would go away if additional variables were added. (Proponents try to anticipate this by showing their results are "robust" to the inclusion of other regressors.) They also suggest alternative functional forms or measures are more appropriate. Worse, they may label results as "meaningless" because of "endogeneity problems" which call for instrumental variables. In response, proponents sometimes assert that x is indeed an exogenous variable and that there is no endogeneity problem. Sometimes this defense is couched in the language of "reduced form regressions." Alternatively, endogeneity concerns may lead to a search for instrumental variables.

It is not surprising to us that these debates invariably generate more heat than light. The problem with using regression models to validate an economic theory is that, absent an economic model that delivers a linear conditional mean specification for y given x, it is impossible to connect regression evidence to deterministic comparative statics predictions. Specifically, for an econometric model to estimate a magnitude of economic interest consistently, the researcher must first use economics and statistics to demonstrate that the relevant economic quantity can be identified using the available data and estimation technique. To see this point more clearly, consider the following example.

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:

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

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 should the researcher prefer? In an informal poll of colleagues, four out of five prefer Model 1 to Model 2. The logic most often given for choosing Model 1 is that it appears to be a cost function. When asked how to interpret the parameters of this regression specification, most say that θ_1 is an estimate of the elasticity of total costs 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 costs is more likely to be "exogenous". To them this means that OLS is more likely to deliver consistent estimates of production or cost parameters.

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 Section 4, we show that Model 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 the Model 1 logarithm of total cost regression and interpret the economic meaning of ordinary least squares parameter estimates. 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 does a researcher 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, $E(y|x)$, which is obtained from $f(x,y)$ as $\int y\, f(y|x)\, dy$. Thus, $\theta_1 = \partial BLP(y \mid x)/\partial x_1$ 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 BLP($y \mid x$) and $E(y \mid x)$, many 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 regression coefficients 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, there is no reason why this need 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 is of limited value. We do not, for example, know how good the linear approximation is. We do not know x causes y, y causes x, or that the coefficients represent a consistent estimate of what would happen to y if we changed x by one unit.

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 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 and z, $E(y \mid x, z)$, where x, y and z are scalars. Suppose that y_d is a customer's demand for electricity during day d, x_d is the price of electricity during day d, and z_d is average temperature during day d. 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 z_d on the right hand-side when we regress y_d on x_d, then we obtain the best linear approximation to $E(y \mid x)$, not $E(y \mid x, z)$. The difference may be very important. For instance, the function $g(x) = E(y \mid x)$ may not depend on x, whereas the function $h(x, z) = E(y \mid x, z)$ may depend on both x and z. In this textbook case, regressing y_d, daily electricity demand, on the daily price x_d could yield a positive or even zero estimated coefficient on x_d, despite the fact that the estimated coefficient on price is large and negative when z_d, the average daily temperature is included.

We anticipate that the point of the previous paragraph may seem obvious to many: omitting a relevant variable can cause bias and produce inconsistent estimates. However, the reasoning here is not as straightforward as the textbook case of omitted variable bias. In the textbook case, the functional form of the conditional expectation is presumed correct (e.g., the conditional mean of y is linear in x and the parameters of interest) and it may be possible to evaluate the bias in OLS coefficients from omitting a relevant regressor. The situation referred to in the previous paragraph adds the complication that we would like to evaluate omitted variable bias when we act as if a linear regression is appropriate when in fact the conditional expectation function is nonlinear. Thus, in addition to the omitted variable, we have to worry that even if we had included the omitted variable in x, that $\partial E(y \mid x)/\partial x \neq \partial BLP(y \mid x)/\partial x$.

Absent a theory that says that y is linearly related to x and z, the effect of omitting a relevant regressor is much harder to evaluate.

3.4 Structural Models, Simultaneous Equations and Reduced Forms

In the remainder of this section, we relate our definition of a structural model to the way the term "structural model" is used in simultaneous equations applications. To begin this discussion, we introduce a textbook linear supply and demand model that is referenced in later discussions of more complex game-theoretic IO models. This model is used to illustrate the conditions a variable must satisfy in order to be a valid instrument in a linear simultaneous equations model. This model also provides a familiar framework to explain why it is not in general possible to consistently estimate the parameters of equations describing the behavior of economic agents from non-experimental data using least squares procedures.

Simultaneous equations models are an important class of structural models in economics because they describe equilibrium phenomena. Simultaneous equations modeling proceeds much as we have described above. The researcher identifies a set of endogenous variables y and exogenous or predetermined variables x. A complete simultaneous equations model contains one equation for each endogenous variable. Each equation either represents the behavior of economic agents, an economic primitive or an economic identity. Simultaneity, however, is not necessarily synonymous with our definition of a structural model. Two examples may help clarify matters.

Example 3

Our first example is the classical linear demand and supply model. Each equation characterizes the behavior of a group of economic agents. The demand curve gives the quantity of the good that consumers would like to purchase at a given price, conditional on other variables thought to affect demand. The supply curve gives how much firms are willing sell at a given output price, conditional on input prices and other supply shifters. The most familiar textbook model is:

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

$$q_t^s = q_t^d, \tag{3}$$

or in matrix notation:

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

$$y_t'\Gamma - x_t'B = \epsilon_t' \tag{4}$$

where Γ and B are matrices containing the unknown parameters that characterize the behavior of consumers and producers, q_t is quantity at time t, p_t is price, y_t is a two-dimensional vector, ϵ_t is a two-dimensional vector of unobserved random variables, and the exogenous x_t consists 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 complete this structural model, 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

$$E(\epsilon_t \mid x_t) = 0 \quad \text{and} \quad E(\epsilon_t\epsilon_t' \mid x_t) = \Sigma. \tag{5}$$

To find out what restrictions the system (3) imposes on the conditional distribution of y given x, we can solve for the endogenous variables as a function of exogenous variables and shocks. Post-multiplying both sides of (4) by Γ^{-1}, and rearranging, gives the reduced form

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

From the conditional moment restrictions imposed on ϵ_t we have

$$E(v_t \mid x_t) = 0, \qquad \text{and} \qquad E(v_t v_t' \mid x_t) = \Omega \tag{7}$$

where

$$\Pi = B\Gamma^{-1}, \qquad v_t' = \epsilon_t'\,\Gamma^{-1}, \qquad \text{and} \qquad \Omega = \Gamma^{-1'}\,\Sigma\,\Gamma^{-1}. \tag{8}$$

From (8), we see that Π and the variance-covariance matrix of the reduced form errors provide information about the structural parameters in Γ. Without restrictions on the elements of Γ, B, and Ω, the only restrictions 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.

Without the economic and stochastic structure that determines *equilibrium* quantities, q_t, and prices, p_t, we cannot deduce the conditional means and variances of q_t and p_t. In other words, 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, as we have argued before, these regression 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 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 $E(q_t \mid x_t)$. ■

To summarize, it is a researcher's economic assumptions about demand and supply that permit her to attach meaning to the coefficients obtained from linear regressions of q_t and p_t on x_t. If we assume the linear supply and demand model in (3) is generating the observed y_t, then the estimates of Π and Ω can be used to recover consistent estimates of the parameters of the stochastic supply and demand equations given in (3). However, if we assume unspecified nonlinear supply and demand equations, then the estimate of Π is only a consistent estimate of the parameters of the best linear predictors of q_t and p_t given x_t. Further, the OLS estimate of Ω now is no longer the variance of one error but two: (1) the difference between the dependent variable and it conditional mean; and, (2) the difference between the conditional mean and its best linear predictor.

In summary, it is economic models that make linear regressions economically meaningful. If we assume stochastic linear supply and demand equations generate y_t, then the equations in (8) 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 no useful empirical content. It only delivers the result that conditional mean of y_t is linear in x_t and the conditional variance is constant. This leads to the question: How do we know that the structural model given in equation (3) is generating the observed y_t?

The answer is hopefully by now familiar: 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.

In defense of structural modeling, one might think that an explicit economic theory is unnecessary if the researcher just "lets the data talk," and performs a purely descrip-

tive analysis. Once again we emphasize that such linear regressions will only produce consistent estimates of best linear predictor functions of the left-hand-side variables given the observed right-hand-side variables. Thus, when a researcher says he would only like to "let the data talk," the data can only speak the language of statistics, informing us about the properties of best linear predictor functions or at best joint and conditional densities.

It is well-known to economic theorists that without assumptions it is impossible 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 this joint density. Unless this 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.

3.4.1 'Magic' Instruments in Simultaneous Equations Models

Econometrics texts are fond of emphasizing the importance of exclusion restrictions for identification. Yet in applied work, most researchers think of simultaneous equations identification problems as one of inclusion – What instruments should I use for my right hand side endogenous variables?

This difference usually arises when applied researchers are initially unwilling or unable to specify all the equations in their simultaneous equations system. This incompleteness in the econometric model reflects an incompleteness in the economic model. This incompleteness can and should raise doubts about instruments. To see why, suppose economic theory delivers the following linear simultaneous equations model

$$\begin{aligned} y_1 &= \beta y_2 + x_1\gamma + \epsilon_1 \\ y_2 &= x_1\pi_{21} + \epsilon_2 \end{aligned} \tag{9}$$

where the ϵ's are independent identically distributed contemporaneously correlated errors and x_1 is an 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 wisdom: "find 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 do this by culling their data sets for variables that might reasonably be viewed as satisfying this criterion.

Is this purely statistical approach valid? The following parable suggests why it is not. We imagine a research assistant who, in an effort to find instruments for the first equation, hits upon the following creative idea. They instruct a computer to create an instrumental variable, x_2, as the sum of x_1 plus computer-generated independent identically distributed random noise. The noise is generated independently of the model errors and x_1 (i.e., they set $x_2 = x_1 + \eta$, where η is independent of ϵ_1, ϵ_2 and x_1). This new variable satisfies the statistical criteria to be a valid instrument: it is by construction uncorrelated with the structural errors and yet correlated with y_2. Thus, it would appear that the research assistant has hit upon a method whereby they could always identify the coefficients in the first equation as long as they initially had at least one exogenous variable and a good random number generator. No economics is required!

We hope that the reader's instincts are that something is amiss here. To see what, recall that it is the matrix of reduced form coefficients Π and the variance of the reduced form errors Ω that we must use to recover the structural parameters. By "finding" x_2 it is as though we have added another variable to the second equation (which already is in reduced form)

$$\begin{aligned} y_1 &= \beta y_2 + x_1\gamma + \epsilon_1 \\ y_2 &= x_1\pi_{21} + x_2\pi_{22} + \epsilon_2 \end{aligned} \tag{10}$$

We now appear to have identification because we have an exogenous variable (x_2) that predicts another endogenous variable in the system (y_2) that is excluded from the equation of interest.[3] The problem with this logic is that economic theory and

[3]More formally, we would estimate the reduced form

$$\begin{aligned} y_1 &= x_1\pi_{11} + x_2\pi_{12} + v_1 \\ y_2 &= x_1\pi_{21} + x_2\pi_{22} + v_2 \end{aligned} \tag{11}$$

and use the four reduced form coefficients to obtain consistent estimates of the four structural parameters: β, γ, π_{21} and π_{22}.

common sense tell us that x_2 does not enter the reduced form. Put another way, the population value of π_{22} is zero! Thus, our computer-generated instrument does not help us (asymptotically) to identify the structural coefficients in the first equation.

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

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

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:

$$\begin{bmatrix} M_2\pi_{21} & M_2 \\ M_2\pi_{21} & M_2 \end{bmatrix}, \tag{12}$$

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 (12) 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 the computer generated instrument example is extreme, but 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 (12).

The value of economic theory is that it provides a defense for why the reduced form coefficient on a prospective instrument is not zero. The statistical advice that led to

computer-generated instruments and x_1^2 does not do this.[4]

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 argue that because these variables are economically related, we do not need a complete simultaneous equations model. The following example discusses this possibility.

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 that market 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, they estimate the regression

$$p_i = POP_i\,\theta_1 + COMP_i\,\theta_2 + \epsilon_i. \tag{13}$$

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

The researcher might, however, claim that equation (13) 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 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 (13).

[4] An element of x_t is a valid instrument in linear simultaneous equations model if it satisfies the conditional moment restrictions (5), $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} x_t\,x_t' = Q$, where Q is a positive definite matrix, and it enters at least one of the equations of the structural model. Our computer generated instrument failed this last requirement.

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. However, as this example shows, the researcher needs to do more. Specifically, to be convincing, the researcher must do two more things. First, the researcher has to explain why it makes sense to *exclude* average income from equation (13). To do this, the researcher will have to provide a more complete economic justification for equation (13). What type of equilibrium relationship does equation (13) characterize? Why is the demand variable POP_i in this equation but average income, which also might be considered a demand variable, not? Second, the researcher also will have to make a case that Y_i enters the reduced form for $COMP_i$ with a non-zero coefficient, or else the rank condition for identification will fail by the logic presented in Example 3. 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 meaningful economic structure.

Example 5

The 1960's and 1970's IO literature contains many studies which regressed firm or industry profit rates ("performance") on market concentration measures ("market structure"). In the late 1960's and early 1970's, 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:

$$\begin{aligned} PROFIT &= \beta_0 + \beta_1 CONC + x_1\beta_2 + \epsilon_1 \\ CONC &= \alpha_0 + \alpha_1 PROFIT + x_2\alpha_2 + \epsilon_2 \end{aligned} \tag{14}$$

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. That is, like inclusion and exclusion restrictions, the decisions about which elements of Γ to normalize to one are decisions that need to come from economic theory if they are to have a economic interpretation. ■

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 macro models are result of 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 restriction 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.

3.4.2 The Role of Non-Experimental Data in Structural Modeling

Virtually all data used in empirical economic research comes from non-experimental settings. The use of non-experimental data can raise significant additional modeling issues for descriptive and structural modelers. In descriptive models, the use of non-experimental data usually raises purely statistical issues. For instance, a researcher may want to describe the relationship between the prices of firms subject to a price cap and the number of competitors. The most general approach to describing this relationship would be to estimate flexibly the joint distribution of prices and competitors. Provided the cap is binding for some firms, the researcher would obtain a density that had a spike of observations at the 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 provided the error term has a conditional mean of zero. In this case, however, the presence of firms at the cap, particularly when there are few competitors, may lead the researcher to question the conditional mean of zero assumption. In response, the researcher might instead use a Tobit or other limited dependent variable formulation of the linear regression model to account for the fact that prices cannot exceed the cap.

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 non-experimental data, let us return to the simultaneous equations demand and supply model for prices and quantities. Suppose the researcher observes price, quantity consumers demand at that price and consumer income (x_1), and 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 the quantity demanded will not equal the quantity supplied at a randomly chosen price. This is because the observed price is experimentally (randomly) determined and it will not in general be the price that equates the quantity demanded with the quantity supplied.

This discussion highlights the importance of the third equation in the demand and supply system (3). In principle, with non experimental data a researcher could observe how much demanders demand at an observed price p and how much suppliers would supply – and these might be different amounts. In practice, researchers typically do not observe the quantity demanded and supplied. They might for example only observed the number of apartments in a city. If there are no regulations restricting supply and demand, then the structural modeler might reasonably account for the non-experimental nature of the price data by assuming that the observed prices are

such that they equate demand and supply. In other words, the researcher might impose the structural demand equals supply equation (3).

How does this non-experimental modeling of price determination compare to the experimental case? One way is view the non-experimental 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 quantities supplied exactly equaling the quantities demanded. This nonrandom sample selection criterion would yield a data set with significantly less information and, more importantly, nonrandom prices. Thus, even though the original data came from an experiment, OLS would 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 non-experimental contexts because they force the researcher to grapple directly with non-experimental 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 non-experimental data – we assumed that the data came from cities where there were no rent floors or ceilings. Had there been rent floors or ceilings, this would change the third equation in our econometric model. For example, with binding rent ceilings, we might assume that the quantity we observe is 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 are reviewed in Madalla (19XX).

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. This process of combining economics and statistics is by no means formulaic. We do, however, believe that there are certain general guidelines and procedures one can follow. In this section,

we propose a framework for constructing and evaluating structural econometric IO models. This framework provides a lens through which we can view 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 models of economic behavior 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 of 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), page xv]

Following in this tradition, we describe a procedure for structural economic modeling that contains three basic steps. The first step is 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 statistical 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
   (c) The information available to actors
2. A list of primitives, including:
   (a) Technologies (e.g., production sets)
   (b) Preferences (e.g., utility functions)
   (c) Endowments (e.g., assets)
3. Variables exogenous to agents and the economic environment, including:
   (a) Constraints on agents' behavior
   (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; quantity demanded
   (b) Profit maximization by firms; quantity supplied
5. An equilibrium solution concept, such as:
   (a) Walrasian equilibrium with price-taking behavior by consumers
   (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
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?

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. Similarly in Example 2, because the researcher did not observe how demand intercepts differed across markets, they were forced to model how the differences in the demand intercepts would impact firms' output decisions.

In general, situations in which economic agents base their decisions on something the researcher does not observed are termed cases of unobserved heterogeneity.

Of course researchers and economic agents can share uncertainty about the economic environment under study. For example, in Example 2, the firm could also be uncertain about the demand intercept. This case differs from the case of pure unobserved heterogeneity in that now firms' quantity decisions are based on their expected (as opposed to realized) demand.[5] A structural auction model could have a similar flavor. 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, depending 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 5

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 consistently estimation α and β in the Cobb-Douglas production function:

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

Since 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 their make 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. In this case, profits equal:

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

[5] We presume that upon learning the demand intercept the firm cannot instantaneously adjust output to maximize profits.

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 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:

$$TC_i = C_0 \; p_{Ki}^{\gamma} \; p_{Li}^{1-\gamma} \; Q_i^{\delta} \; A_i^{-\delta}, \tag{15}$$

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:

$$\ln TC_i = \ln C_0 + \gamma \ln p_{Ki} + (1-\gamma) \ln p_{Li} + \delta \ln Q_i - \delta \ln A_i \tag{16}$$

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:

$$\ln TC_i = C_1 + \gamma \ln p_{Ki} + (1-\gamma) \ln p_{Li} + \delta \ln Q_i - \delta \ln u_i. \tag{17}$$

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.

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, the firms could have different total costs. What is left to explain, is how a researcher would estimate the production parameters. This is a non-trivial 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 the firm i's expected profit function is:

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

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:

$$E\left[\,\pi(p_i^{er} K_i, L_i)\right] = p_i^{er}\, E\left[\,A_i L_i^{\alpha} K_i^{\beta}\right] - p_{Ki} K_i - p_{Li} L_i. \tag{19}$$

The first-order conditions for expected profit maximization imply

$$L_i = \left[\frac{\alpha\, p_{Ki}}{\beta\, p_{Li}}\right]\, K_i \tag{20}$$

Observed total costs therefore equal

$$TC_i = \frac{\alpha + \beta}{\beta}\; p_{Ki}\; K_i \tag{21}$$

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

$$Q_i^a = D_0\; TC_i^{\alpha+\beta}\; p_{Ki}^{-\beta}\; p_{Li}^{-\alpha}\; A_i \tag{22}$$

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

$$\ln\, Q_i^a = \ln\, D_0 + (\alpha + \beta)\, \ln\, TC_i - \beta\, \ln\, p_{Ki} - \alpha\, \ln\, p_{Li} + \ln\, A_i. \tag{23}$$

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

$$\ln Q_i = D_1 - (\alpha + \beta) \ln TC_i - \beta \ln p_{Ki} - \alpha \ln p_{Li} + \eta_i. \tag{24}$$

where $\eta_i = \ln A_i - E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln TC_i]$. The constant term $D_1 = \ln D_0 + E[\ln A_i \mid \ln p_{Ki}, \ln p_{Li}, \ln Q_i]$ absorbs the nonzero conditional mean of the efficiency differences. 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 (17) 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.

We now can better understand 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 reseacher's data were generated. Since no economic model is perfect in practice, the researcher often will be left choosing among imperfect assumptions and models. No statistical test can automate the process of choosing among models. 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 6

This example narrowly interprets optimization errors as the failure of agents' decisions to satisfy first-order necessary conditions for optimal decisions exactly. 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:

$$\min_{\lambda \geq 0} \left[\max_{x \geq 0} U(x, \eta) + \lambda (M - p'x), \right] \tag{25}$$

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:

$$\frac{\partial U}{\partial x_i}(x_i, \eta_i) - \lambda_i \, p_i = 0. \tag{26}$$

These equations yield the $i = 1, ..., n$ 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 (26). Specifically, suppose

$$\frac{\partial U}{\partial x_i}(x, \eta) - \lambda \, p_i \, \nu_i = 0. \tag{27}$$

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 demands 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 7

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.

Following equation (27), we might assume each consumer has the utility function is $U(x,y) = x^a\ y^b$. Solving the first-order conditions we obtain

$$\frac{a}{x} = \lambda\, p_x\ \nu_{xi}, \qquad \frac{b}{y} = \lambda\, p_y\ \nu_{yi}, \qquad p_x\, x + p_y\, y = M, \tag{28}$$

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

$$\lambda = \frac{\alpha_i + \beta_i}{M} \qquad \text{with} \qquad \alpha_i = \frac{a}{\nu_{xi}} \qquad \text{and} \qquad \beta_i = \frac{b}{\nu_{yi}}, \tag{29}$$

$$x = \frac{\alpha_i}{\alpha_i + \beta_i}\ \frac{M}{p_x} \qquad \text{and} \qquad y = \frac{\beta_i}{\alpha_i + \beta_i}\ \frac{M}{p_y}. \tag{30}$$

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 intertemporarlly, 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 according to standard producer theory.

Example 7

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:

$$p_L\,\nu_L = \lambda\,\alpha\,K^{\beta}\,L^{\alpha-1}, \qquad p_K\,\nu_K = \lambda\,\beta\,K^{\beta-1}\,L^{\alpha} \qquad \text{and} \qquad Q = K^{\beta}\,L^{\alpha}, \tag{31}$$

where λ is the Lagrange multiplier associated with the constraint that the firm produce using the production function, and ν_{Li} and v_{Ki} are unit mean, positive random variables representing optimization errors. Solving these three equations yields following two factor demands:

$$L = Q^{\frac{1}{\alpha+\beta}}\,\left[\frac{p_K}{p_L}\right]^{\frac{\alpha}{\alpha+\beta}}\,\left[\frac{\beta\,\nu_K}{\alpha\,\nu_L}\right]^{\frac{\alpha}{\alpha+\beta}} \tag{32}$$

$$K = Q^{\frac{1}{\alpha+\beta}}\,\left[\frac{p_K}{p_L}\right]^{-\frac{\beta}{\alpha+\beta}}\,\left[\frac{\beta\nu_K}{\alpha\,\nu_L}\right]^{\frac{-\beta}{\alpha+\beta}}. \tag{33}$$

An implication of the optimization errors, ν_{xi} and ν_{yi}, is that the symmetry restriction implied by cost-minimization behavior fails. Specifically, the restriction

$$\frac{\partial L}{\partial p_K} = \frac{\partial K}{\partial p_L} \tag{34}$$

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 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 (27) are allowed to take on values less than zero, for certain realizations of ν 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 because of data reporting and collection errors. 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 often is little prior information about the properties of the measurement error. This means that the researcher typically 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 strong spike at ten units.

Measurement error also is 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 apply 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
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 tradeoff between data and parametric flexibility. Larger datasets usually allow greater parametric flexibility. A second criterion is that the functional form should permit flexibility when estimating economic quantities of interest. 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 expected 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

$$\frac{\partial \pi(y, x, \epsilon, \beta)}{\partial y} = 0. \tag{35}$$

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

$$p(y \mid x, \theta, \beta) = f(h^{-1}(y, x, \beta), \theta) \left| \frac{\partial h^{-1}(y, x, \beta)}{\partial y} \right| \tag{36}$$

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 tradeoff between efficiency and robustness. Maximum likelihood techniques may be inconsistent if not all of the distributional assumptions hold. 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 be what identifies economic structure.

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 and statistical 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 Supply 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 complex 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 otherwise exercising excessive market power.

Most researchers studying competition do not know when firms are competing or colluding. Instead, they 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 they 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 heavily 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 non-experimental price and quantity data to identify the extent of industry competition.

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 markup price above cost. The IO literature typically draws a distinction between demand models for homogeneous product markets and differentiated product markets. 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:

$$\text{Market Demand} \qquad Q = h(P, Z, \beta, \nu) \tag{37}$$

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 competition, however, 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. Once these are specified, the researcher must then introduce assumptions about how firms interact. These assumptions about firms'

strategic interactions affect the structure of the first-order conditions that characterize firms' optimizing 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:

$$\text{"Industry Supply"} \qquad P = g(Q, W, \theta, \eta) \tag{38}$$

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 a error term for market supply.

Equations (37) and (38) 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 a competitive supply equation 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 (37) and (38). Even the researcher is willing to make distributional assumptions about the joint density of ν and η, without assumptions on the functional form of (37) and (38), 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 (1972) first estimated the extent of 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 U.S. railroad cartels is one of the first papers in IO

to devise a sophisticated structural econometric model of a cartelized industry.[6] The economic logic for Porter's empirical model comes from Green and Porter (1983). 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 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:

$$\ln Q_t = \alpha + \epsilon \ln P_t + Z_t \gamma + \nu_t, \tag{39}$$

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

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

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

$$C_i(q_{it}) = a_i \; q_{it}^{\delta} \tag{40}$$

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 a firm-level "supply" of output curve for a range models of competition that can be aggregated and to obtain an industry-level "supply" curve.

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, but this 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, 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, 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:

$$p_{it} \; \left(1 + \frac{\theta_{it}}{\epsilon}\right) = a_i \; \delta \; q_{it}^{\delta - 1} \tag{41}$$

Here,

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

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 response to a one unit change in firm i's output level.

Many economists argue that it is impossible to interpret the conjectural variation parameter independent of its value. 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, the firm believes that any change in its output will be met with no change in the output of its competitors, so that $\frac{\partial Q_{-it}}{\partial q_{it}} = 0$, 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 perceives 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 (38). 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 (41) by its quantity,

$$p_t \left[1 + \frac{\theta_t}{\epsilon}\right] = D\, Q_t^{\delta - 1}, \tag{42}$$

where

$$D = \delta \left(\sum_{i=1}^{N} a_i^{1/(1-\delta)} \right)^{1-\delta}, \tag{43}$$

$$\theta_t = \sum_{i=1}^{N} s_{it}\, \theta_{it}, \tag{44}$$

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:

$$\begin{array}{rcll} \ln Q_t - \epsilon \ln p_t & = & \alpha + Z_t\gamma + \nu_t & \text{Demand Equation} \\ -(\delta - 1)\ln(Q_t) + \ln p_t & = & \lambda + \beta I_t + W_t\phi + \eta_t & \text{Supply Equation} \end{array} \tag{45}$$

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 that 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. Were it otherwise, the rail firms would have optimized against the distribution of these errors rather than the errors themselves and the endogenous variables of the model would no longer be functions of the realized errors. Porter also assumes the demand and supply errors are uncorrelated with the right-hand-side exogenous variables. By inspection of the aggregated first-order conditions for profit-maximization in equation (42), 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 non-zero 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 consistent with Porter's formulation, the nonzero average error is absorbed into the supply equation's constant term, giving $\lambda + E(\, I_t) = \lambda + \beta\tau$. 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 and demand model (3), discussed in the previous section. Porter's demand and supply system has the form

$$y_t'\Gamma + x_t'B = E_t' \tag{46}$$

$$[\ln Q_t \ \ln p_t] \begin{bmatrix} 1 & -(1-\delta) \\ -\epsilon & 1 \end{bmatrix} + [\, 1 \ Z_t' \ W_t' \,] \begin{bmatrix} -\alpha & -(\lambda+\beta\tau) \\ -\gamma & 0 \\ 0 & -\phi \end{bmatrix} = [\, \nu_t,\ \beta\,(I_t - \tau) + \eta_t]$$

At this point, 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 $E(E_t E_t') = \Sigma$ in equation (46) by three-stage least squares.

Notice, however, that absent 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 (iid) Bernoulli process, independently distributed of the normally distributed demand and supply errors.

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 potential asymmetry 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 non-normality 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 price and quantity $y_t = (\ln p_t, \ln Q_t')$ 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 revisionary regimes. However, no amount of data will allow the reseacher to distinguish between this regime-switching structural model and a conventional linear simultaneous equations model with no regime switching. Specifically, suppose E_t has the density

$$\begin{aligned} g(E_t) = \quad & \tau \tfrac{1}{2\pi}\, |\Sigma|^{-1/2} \exp(-\tfrac{F_{1t}'\Sigma^{-1}F_{1t}}{2}) \\ & +(1-\tau)\tfrac{1}{2\pi}\, |\Sigma|^{-1/2} \exp(-\tfrac{F_{2t}'\Sigma^{-1}F_{2t}}{2}) \end{aligned}$$

where

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

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. 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 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 affect results. This example also illustrates our earlier meta-theorem that: absent assumptions about the economic model generating the observed data, the researcher can not do much beyond describing 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:

$$y_t'\Gamma = x_t'D + I_t\Delta + U_t' \tag{47}$$

where

$$\Gamma = \begin{bmatrix} 1 & -(1-\delta) \\ -\epsilon & 1 \end{bmatrix}, \quad \Delta = \begin{pmatrix} 0 \\ \beta \end{pmatrix}, D = \begin{bmatrix} \alpha & \lambda \\ \gamma & 0 \\ 0 & \phi \end{bmatrix}, U_t = \begin{bmatrix} \nu_t \\ \eta_t \end{bmatrix}, \quad \text{and} \quad U_t \sim N(0, \Sigma). \tag{48}$$

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

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

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

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

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 $f^{true}(y_t, x_t)$, from which can derive, the conditional density of y_t given x_t. The fact that $f^{true}(y_t|x_t)$, the true conditional density, can be factored into the product of two conditional normal densities of y_t given x_t and I_t 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 for the underlying economic interaction.

Without imposing this economic structure on $f(y_t|x_t)$, 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 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 $f^{true}(y_t, x_t)$ using the techniques in Silverman (1986). From this estimate, the researcher could compute an estimate of $E(y_t|x_t)$ 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

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

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 non-normal 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 zero and 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.

This identification problem is common to all industrial organization models of firm and industry behavior models, as we shall see.

The strength of Porter's model is that it both identifies 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 a nice 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. It would be difficult to imagine how one could rationalize this variation with a descriptive model of prices and quantities.

6 Market Power Models More Generally

Porter's model is an example of IO models that draw inferences about competition from data on market-clearing prices and outputs. Because these are among the most widely used empirical models in industrial organization, it is worth going beyond Porter's model to consider strategies that other studies have used 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 equate competition with price equal to marginal cost. When price is above marginal cost, firms are said to have "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 price and quantity. 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 inferences about the extent of competition. That is, they draw inferences about price-cost

margins without acknowledging that their inferences depend critically on their economic assumptions, or even an explicit statement of what the economic assumptions are underlying the interpretation of a vector of estimated parameters. As we have shown in a number of previous examples, economic assumptions typically can and will alter inferences one might draw about the extent of competition. The debate about whether the estimates are believable then hinges on the plausibility of maintained assumptions or, as in Porter's case, the availability of alternative evidence.

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 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 or markets where firms choose supply schedules (prices and quantities).

6.1 Estimating Price-Cost Margins

Since the late 1970's, 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:

$$P + \theta_i \, q_i \, \frac{\partial P}{\partial Q} = MC_i(q_i), \tag{49}$$

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 (49), we obtain

$$\theta_i = \frac{P - MC_i(q_i)}{-\, q_i \, \frac{\partial P}{\partial Q}} = \frac{P - MC_i(q_i)}{P} \frac{1}{\alpha_i \, \epsilon}. \tag{50}$$

From this equation, we see that θ_i provides essentially the same *descriptive* informa-

tion 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 (50) 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 (50) poses a more difficult estimation problem. Equation (50) 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 function 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. The usual case in practice is that we have little reason to prefer one parametric marginal cost or demand specification to another.

Despite the difficulty of inferring marginal costs from price and quantity data alone, many studies go further, choosing to interpret θ_i as a measure of firm behavior. To understand where this behavioral interpretation comes from, we return to the economic rationale underlying equation (49). In equation (49), θ_i is a placeholder for the derivative:

$$\theta_i \; = \; \frac{d\,Q}{d\,q_i}, . \tag{51}$$

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 (49) 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 since estimates of θ_i are predicated on a specific functional form for marginal costs.

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 iden-

tical, 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. Only the descriptive interpretation makes much sense to us. There are, however, a few instances in which θ_i sensibly corresponds to a specific model of firm behavior. The leading case is perfect 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 (49) and using them to estimate the parameters of demand and cost functions. One can then use non-nested tests, as in Bresnahan (1987), to compare these different models of behavior.

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 (49), 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 $\theta_i = 1.7$? 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.[7] 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 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 following equation

$$P + \theta Q \frac{\partial P}{\partial Q} = MC(Q) \tag{52}$$

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

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

cost curve. He used observations on market-clearing prices and quantities from a cross-section of U.S. 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 (52) 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 (52) 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 (49) across firms to obtain the relation

$$P + \frac{\partial P}{\partial Q} \sum_{i=1}^{N} \frac{\theta_i \, q_i}{N} = \sum_{i=1}^{N} \frac{MC(q_i)}{N} \tag{53}$$

which we can rewrite as

$$P + \theta \frac{\partial P}{\partial Q} Q = \overline{MC(q_i)}. \tag{54}$$

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 margin parameters, and $\overline{MC(q_i)}$ is the average of the N firms' marginal costs. While this equation "looks" like the industry aggregate equation (52) 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 (52) 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 demand curves and the same marginal revenues. This is a non-trivial restriction which would require justification on a case by case basis. A number of studies, beginning with Gollop and Roberts (1979), Appelbaum (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 (52), we see that the notation $MC(Q)$ 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 (49). This obviously cannot generally be true. Equation (54) shows why this is so. To explore this point further, it is useful to assume that firms have linear marginal costs of the form:

$$MC(q_i) \;=\; c_{0i} \;+\; c_{1i}\, q_i \tag{55}$$

In this case, we can rewrite equation (54) as

$$P \;+\; \tilde{\theta}\, Q\, \frac{\partial P}{\partial Q} \;=\; \bar{c_o} \;+\; \bar{c_1}\, Q \;+\; \psi \tag{56}$$

where

$$\tilde{\theta} \;=\; \frac{\sum_{i=1}^{N} \frac{\theta_i}{N}}{N} \tag{57}$$

$$\bar{c_0} \;=\; \frac{1}{N} \sum_{i=1}^{N} c_{0i} \qquad\qquad \bar{c_1} \;=\; \frac{1}{N} \sum_{i=1}^{N} c_{1i} \tag{58}$$

$$\psi \;=\; Cov(c_{1i}, q_i) \;-\; Cov(\theta_i, q_i)\, \frac{\partial P}{\partial Q} \tag{59}$$

and $Cov(x, y)$ equals the covariance (calculated over firms in the industry) between x and y. If the ψ term is zero, then equation (54) and equation (52) 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 (52) as having an error term that includes ψ . To the extent that 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 effect 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 (52). Since 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 $MC_i = c_0 + c_1 q_i + c_2 w_i$, 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

$$\begin{array}{rcl} Q & = & D\,(\,P,\,Y,\,\alpha\,) \\ P & = & c_0 \,+\, (\,c_1 \,+\, \frac{\partial D^{-1}}{\partial Q}\,\theta_i\,)\,q_i \,+\, c_2\,w_i \end{array} \tag{60}$$

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:

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

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. 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 to obtain the supply equation

$$P \,=\, c_0 \,+\, \left(\, c_1 \,+\, \frac{\theta_i}{\alpha_1 \,+\, \alpha_2\, Y} \,\right)\; q_i \,+\, c_2\, w_i \tag{62}$$

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. Returning to the specification in equation (54), 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 we assumed that the firms had the same 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:

$$P = \alpha - \beta Q^{1/\gamma} \tag{63}$$

This inverse demand function leads to the direct estimator (equation (50) above)

$$\theta^1 = -\gamma \frac{P - c}{\alpha - P} \tag{64}$$

which illustrates how the demand parameters affect the direct estimate. This inverse demand function also yields a transformed equation (52)

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

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 (20) gives the equation

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

This equation makes it clear that absent further assumptions, we cannot identify θ if it is a constant parameter. One way around this problem is to recognize from equation (54) 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 (20) 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 well draw poor estimates of θ if in fact $\gamma \neq 1$.

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 margins and the appropriate model of competition.

7 Models of Competition in Differentiated-Product Markets

The previous two sections discussed how IO economists have used price and quantity data to draw inferences about strategic interactions among 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, IO economists 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 unobserved demand, costs and competition from just data on prices and quantities.

7.1 Neoclassical Demand Models

In the late 1980s and 1990s, empirical IO economists began to focus on modeling competition in differentiated products markets such as cars, computers and break-

fast cereals. 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 models for homogeneous products is that the researcher must specify separate "demand" and "supply" equations for each product. Thus, instead of working with two-equation systems such as (37) and (38), the researcher has a J-product demand system:

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

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

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

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' demand and costs, the researcher will be unable to draw inferences about firms' markups or competition. This, again, is because the researcher does not have information about costs and competition. Thus, 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 some demand and cost variables W and Z. Only by making functional form assumptions for demand and costs, and making assumptions about unobserved strategic interactions, can the researcher recover information about demand and cost primitives. This means that we again have to consider how specific functional form and other assumptions affect a researcher's ability to use price and quantity data to draw inferences about markups and competition.

The main new issue posed by differentiated products is one of scale. In particular, now the researcher has to specify a set of demand functions – potentially involving dozens or hundreds of products. Absent restrictions on demand and firm behavior, the number of parameters in these equations easily outnumbers the number of observations in conventional market-level price and quantity datasets. This problem has led IO researchers to focus on the problem of how best to formulate parsimonius, 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 US ready-to-eat breakfast cereal makers. Absent cost data, inferences about price-cost margins and competition have to drawn from the prices and quantities of different cereal

brands. As there are over 50 major brands of cereals, a researcher has to formulate a 100 equation structural equation model. Each equation of the model conceivably could contain dozens of parameters. For instance, paralleling Porter's homogeneous specification, if the researcher adopted the log-linear demand system:

$$\begin{array}{rcl}
\ln Q_1 & = & \beta_{10} + \beta_{11} \ln y + \beta_{12} \ln P_1 + \beta_{13} \ln P_2 + ... + \beta_{1,50} \ln P_{50} + Z_1\gamma_1 + \epsilon_1 \\
\ln Q_2 & = & \beta_{20} + \beta_{21} \ln y + \beta_{22} \ln P_1 + \beta_{23} \ln P_2 + ... + \beta_{2,50} \ln P_{50} + Z_2\gamma_2 + \epsilon_2 \\
\vdots & = & \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\
\ln Q_{50} & = & \beta_{50,0} + \beta_{50,1} \ln y + \beta_{50,2} \ln P_1 + \beta_{50,3} \ln P_2 + ... + \beta_{50,,50} \ln P_{50} + Z_{50}\gamma_{50} + \epsilon_{50},
\end{array} \tag{67}$$

they would have to estimate at least 2,600 parameters! Such unrestricted parameterizations easily exceed the number of observations obtainable from public sources.

The scale of differentiated product models also raises significant econometric and computational challenges. When the equations in (65) and (66) are nonlinear in the demand and cost errors, it is difficult to devise computationally convenient and consistent estimators. 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 can present nontrivial computational issues. For instance, the system need not have a unique solution, or any real-valued solution, for all error and parameter values. Although these complications can sometimes be dealt with in estimation, they may still reappear should the researcher wish to perform counterfactual calculations. For instance, there may be no real-valued prices that solve (66) for a particular counterfactual.

Both the scale of these systems and these econometric issues have prompted IO researchers to look for ways to simplify differentiated product models. Initial efforts focused on trying to simplify traditional Marshallian demand systems as a way of limiting parameters. Many early simplifications relied on ad hoc parameter restrictions or the aggregation of products.[8] For example, to estimate (67) a researcher might constrain a product's cross-price elasticities to all be the same.[9] Assumptions such as this, while computationally convenient, can have unattractive economic consequences, as the own-price and cross-price elasticities of demand enter the mark-up equations (66) that determine how far price is above marginal cost. Such ad hoc price elasticity restrictions can result in the serious misestimation of price-cost margins.

Multi-level demand specifications provide a somewhat more flexible method for restricting demand parameterizations.[10] In a multi-level demand specification, the re-

[8] Bresnahan's (1989) section 4 reviews early efforts. Deaton and Muelbauer (1980) provide a survey of neoclassical demand models.

[9] 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).

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

searcher 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 Chocola, etc.).

Although multi-stage models also restrict price elasticities, they permit flexible cross-price elasticities for products within a particular product category. For example, the researcher can estimate a flexible Marshallian 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 consumers' overall kids' cereals spending. These indirect price effects are therefore not as flexible as the within category cross-price effects. Whether this lack of flexibility matters much for estimates of price-cost margins, is as yet unclear.[11] A series of theoretical papers beginning with Gorman (1959) have, however, explored the restrictions that multi-stage budgeting models place on consumer preferences, and how these restrictions affect compensated and uncompensated price effects.[12]

Other recent work in the Marshallian demand system tradition has explored reducing the number of demand parameters by working with reduced forms or constraining cross-price effects to depend on estimable functions of covariates.[13] Pinske, Slade and Brett (2000) and Pinske and Slade (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 form of demand reflects the way consumers evaluate product attributes. (See also Davis (2000).)

Besides having to grapple with how best to restrict parameters, each of the above approaches also has to address the endogeneity 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 more nonlinear systems, identification and estimation questions become much more complicated. For example, the implicit "reduced form"

[11] Nevo (1997).

[12] See for example Gorman (1970), Blacorby et al. (1978) and Hausman et al. (1994).

[13] 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.

for the nonlinear (65) and (66) system:

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

may not be available in closed form. (Here $Z = (Z_1, ..., Z_J)$ and $W, \beta; \theta, \nu$ and η are similarly defined collections.) These equations also need not have a solution or a unique solution for all values of the right hand side variables and errors.

Large demand systems also pose difficult computational and estimation issues. For example, system estimation methods, such as full information maximum likelihood applied to (68) may prove infeasible simply because of the number of parameters, Alternatively, single equation methods such as generalized method of moments and nonlinear instrumental variable methods raise efficiency issues. Often, efficiency, and indeed consistency, is tied to the selection of instruments for prices and quantities. The reduced forms (68) suggest many possible instruments, including a product's own attributes and cost variables, and other products' attributes and cost variables. Unfortunately, most IO data sets do not have product-specific or firm-specific cost information. Even when they do, they cannot make use the information because it is extremely highly correlated or collinear. The lack of good cost covariates has forced most researchers to use non-price attributes as instruments, or in some panel data contexts, the prices of products in other markets as instruments.

The use of other prices and non-price attributes as instruments raises delicate identification issues. Consider, for instance, how one might use panel data on prices and quantities in different geographic markets. Several researchers have proposed using prices in nearby markets as instruments for prices. Thus, to estimate cereal demand in San Francisco, the researcher might use contemporaneous cereal prices in Los Angeles as instruments. The key modeling question here is the same as in Section 4: How do we know prices in other markets are valid instruments? The answer again has to come from economics.

Hausman (1996) uses economics to motivate his use of nearby price instruments to model cereal demand. He assumes that the price for brand j in market m has the form

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

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

markups. Although he does not spell out the underlying economic model that delivers this equation, it clearly parallels the first-order conditions for profit maximization found in equations such as (49).

For prices in market n to be valid instruments for prices in market m, they must be correlated with prices in market m and uncorrelated with the demand error for product j in market m. Because the α's represent unobserved product and market-specific factors that affect mark ups they could well be correlated with the demand errors. As an example, one could imagine San Franciscans' unobserved health conscious attitudes leading them to have a higher demand for organic cereals and, as a result, them paying higher markups on average. Hausman deals with this concern by removing the brand-market α's using product-market fixed effects. Letting $\tilde{}$ denote the residual prices from these regressions, his adjusted prices have the form:

$$\widetilde{\ln p_{jnt}} \quad = \quad \delta_j \, \widetilde{\ln c_{jt}} + \widetilde{\nu_{jnt}}. \tag{69}$$

That is, they contain only adjusted national marginal costs and residual cost and demand factors affecting markups. At this point, Hausman makes two critical assumptions: (1) the adjusted time-varying national marginal costs $\ln c_{jt}$ are uncorrelated with the demand errors in other cities; and (2) the residual demand and cost factors affecting markups are uncorrelated with the demand errors in other cities.

How likely is it that these two assumptions are satisfied? Economics can provide no general answer here. Indeed, these issues have been vigorously debated by Hausman and Bresnahan (1997). The advantage of Hausman's model for prices is it helps focus the debate on the institutional features of the market that might make his assumptions valid. For example, one could criticize Hausman's assumptions by saying that there are common national unobserved seasonal factors that affect both the demand and marginal cost of cereal brands. Such factors would invalidate assumption (1), unless one included (as Hausman did) monthly dummy variables in the instrument list. Condition (2) could fail for similar reasons, but to know for sure whether this is a problem, one has to use economic arguments to understand when unobserved demand factors might affect the residual mark up terms in $\widetilde{\nu_{jnt}}$. Bresnahan (1996) provides such a model in which a firm's 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.[14] 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

[14]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.

firms under study.

IO economists also have sought to use non-price attributes as instruments for prices and quantities. Similar concerns, however, can be raised about non-price 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 non-price 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 choose departure times. For example, carriers may choose different departure times so as to differentiate their flights and charge higher prices.

If firms set non-price attributes using information unavailable to the econometrician, then we can no longer be certain that product attributes are valid instruments. While in principle we might follow Hausman's example, and use other products' characteristics as instruments, we would, like Hausman, have to develop an economic model of product characteristic choice. Such an approach would significantly complicate a differentiated product model.

In some applications, researchers have defended the use of non-price attributes as demand instruments with the argument that they are "predetermined". Implicit in this defense is the claim that firms find it prohibitively expensive to change non-price attributes once set. As a result, non-price product characteristics can reasonably thought of as being uncorrelated with short run unobserved demand variables that affect prices. 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 the unobserved factors that influenced the manufacturer's initial choice of characteristics do not persist through time. These and other endogeneity problems continue to make the search for identifying assumptions in differentiated product models an active area of research.

7.2 Micro-Data Models

Our discussion of the product-level demand specifications in (65) so far has focused on using product-level price and quantity data to draw inferences about competition and price-cost margins. Product-level demand specifications, however, potentially obscure important differences among consumers that impact firms' price and quantity decisions and competition. For example, the recognition that consumers have different willingnesses to pay immediately raises the question of why firms would not attempt to price discriminate or otherwise segment consumers. Frequent buyer

discounts are a ready example of this. More sophisticated bundling strategies, such as offering options packages on new cars also come to mind. Demand systems that do not model heterogeneity in consumer tastes cannot begin to contemplate these possibilities, or indeed what would happen if a new product were introduced into a market. These possibilities suggest that accounting for consumer heterogeneity ought to be an important part of differentiated product models. Most neoclassical demand system models, however, explicitly adopt a representative consumer framework.

When, as is sometimes the case, researchers have individual-level consumption data, they can model the heterogeneity that firms may perceive in consumers' tastes. Although such demand data ideally would come from the firms making pricing and production decisions, in practice IO researchers have had to rely on government data and non-proprietary marketing surveys. The main modeling issue that researchers face in using such data is how to go from individual-level demand specifications to the product-level demand functions that firms are presumed to use when formulating price and quantity decisions.

Prior research has taken one of two approaches to this problem. The first estimates consumer-level demand models using representative samples of consumers. The researcher then uses sample weights to aggregate consumer demands to product-level estimates of firm demand curves. The second approach does not have information on the distribution of individual consumer tastes. Instead, it estimates the distribution of consumer tastes along with other demand and supply parameters from aggregate price and quantity information.

In what follows, we explore some of the advantages and disadvantages of these two approaches. To focus our discussion, we follow the existing literature and consider discrete demand specifications. Specifically, these models presume that consumers buy at most one unit of one product from among J products offered.[15] While these unitary demand models are literally applicable to only a few products, such as new car purchases, they have been applied to estimate the demands for a range of products.

The distinguishing feature of a discrete choice demand model is that consumer demands are probability statements. Once the researcher adopts a specific probability model for choice, product-level demands simply are sums of consumers' purchase probabilities. These highly nonlinear demand systems have two compensating advantages. First, the discrete choice framework readily allows the researcher to model how product attributes affect consumer tastes and decision making. In particular, consumers preferences over a large number of products now can be reduced to a short list of product attributes. Thus, instead of trying to model 50 cereal products directly, a researcher can reduce the consumers' choice problem to a choice over a

[15]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).

cereal's sweetness, crunchiness and fiber content. A second advantage is that discrete choice models can generate rich substitution patterns for products. These patterns can change depending on the proximities of product attributes.

As with homogeneous-product structural competition models, differentiated product models contain many economic, functional form and stochastic assumptions that will affect the researcher's inferences about demand, consumer preferences and price-cost margins. In what follows we use our framework for building structural models to evaluate two early differentiated product models describing US new-car price-cost margins. The first, by Goldberg (1995), uses representative household new-car purchase data to estimate household-level purchase probabilities for different new car models. She then sums these household-level probabilities to obtain demand estimates at the market level. Assuming these market-level estimates are what firms use when determining prices, she uses them in firms' profit-maximization conditions to draw inferences about products' unobserved marginal costs and new car price-cost margins. 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 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, albeit 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. Consumer i posses the time t conditional indirect utility function:

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

where x_{jt} is a $K \times 1$ vector of non-price attributes of car j (such as size and horsepower), p_{jt} is the car's price, and ω_{ijt} represents consumer-level variables. She assumes that consumer i purchases at most one new or used car. In particular, consumer i chooses new car j provided $U(x_{jt}, p_{jt}, \omega_{ijt}) \geq \max_{k \neq j} U(x_{kt}, p_{kt}, \omega_{ijt}; \theta)$. If firms knew everything about consumers' tastes, they would calculate product demand as

$$\text{Demand for Product } j = \sum_{i=1}^{M_t} I(\ i \text{ buys new car } \ j) \tag{70}$$

where M_t is the number of potential new car buyers at time t and $I(Arg)$ is a zero-one discrete indicator function that is one when Arg is true. Of course, neither Goldberg or the new car manufacturers observe everything relevant to consumers' decisions. Firms instead base pricing decisions on what they expect demand to be:

$$\text{Expected Demand} = q_{jt}^e = \sum_{i=1}^{M_t} E\left(U(x_{jt}, p_{jt}, \omega_{ijt}) \geq \max_{k \neq j} U(x_{kt}, p_{kt}, \omega_{ikt}; \theta) \right). \tag{71}$$

In this expression, E is the firm's expectation over the unobservables in ω_{ijt}. The firm is assumed to know the size of the market M_t. The expectation in (70) can equivalently be expressed as the sum of firms' probability assessments that consumers will buy model j:

$$q_{jt}^e = \sum_{i=1}^{M_t} Pr(\ i \text{ buys new car } \ j). \tag{72}$$

Goldberg of course does not observe firms' expectations. The initial step of her estimation procedure therefore seeks to approximate $Pr(\cdot)$ 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 Statistics Consumer Expenditure Survey. This survey is a stratified random sample of approximately 4,500 to 5,000 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 car attributes.

A critical component of her expected demand model is the list of attributes that enter consumers' utility functions. While economics provides clear guidance that the transactions price is a relevant attribute, economics provides little guidance about the other attributes that 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. The two questions that cannot easily be answered by this approach are whether these are the only attributes consumers care about and how consumers trade off the availability and prices of vehicle options.

In estimating the expected demands faced by new car manufacturers, Goldberg relies on the representativeness and accuracy of Consumer Expenditure Survey. Her assumption that her probability model replicates the firms' assessments of consumer behavior allows her to replace $Pr(i \text{ buys new car } j)$ in (72) with an econometric estimate, $\hat{Pr}(k \text{ buys new car } j)$ – sample household k's purchase probability. The assumption that the CES sample is representative of the M_t consumers permits her to replace the sum over consumers in (71) with a weighted sum of the estimated household purchase probabilities: (73) by

$$\text{Estimated Demand for Product } j = \sum_{k=1}^{S_t} w_{kt}\, \hat{Pr}(\ k \text{ buys new car } j) \qquad (73)$$

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. Instead, Goldberg uses manufacturers' suggested retail ("list") prices to construct proxies for the unobserved wholesale prices. Specifically, she assumes that the manufacturer's wholesale price is a constant percentage of a new car's suggested list price. While this assumption facilitates estimation, it is unclear exactly what behavior leads manufacturers to set suggested list prices in this way.[16]

Goldberg models manufacturers' decisions about wholesale prices as outcomes of a static Bertrand-Nash pricing game in which manufacturers maximize expected U.S. profits. The expectation in profits is taken over the demand uncertainty in each ω_{ijt}.[17]

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

[17] In principle, the firm also might be uncertain about its marginal cost of production. Goldberg

Thus, firm f maximizes

$$\max_{p_{jt}^W} \sum_{i=1}^{n_{ft}} (p_{it}^W - c_{it})\, E(q_{it}) \tag{74}$$

where p^W is wholesale price, n_{ft} is the number of new car models offered by firm f and the c_{it} are constant marginal production costs.

The first-order conditions that characterize manufacturers' wholesale pricing decisions have the form:

$$p_{jt}^W\, q_{jt}^e + \sum_{i=1}^{n_{ft}} \frac{p_{it}^W - c_{it}}{p_{it}^W} q_{it}^e\, \epsilon_{ijt} = 0 \tag{75}$$

where $q_{it}^e = E(q_{it})$, and $\epsilon_{ijt} = \frac{p_{jt}^W}{q_{it}^e} \frac{\partial q_{it}^e}{\partial p_{jt}^W}$ 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. In practice, Goldberg's supply side first-order conditions are further complicated by the reality that Japanese manufacturers during the early and mid-1980s faced voluntary export restraints. Goldberg models these voluntary restraints as potentially binding constraints on firms' quantities. Their introduction considerably complicates matters because Goldberg must worry that they may cause there to be no pure-strategy pricing equilibrium. In an attempt to overcome this problem, she assumes that excess demand for Japanese imports would not generate extra demand for domestic vehicles. She claims this assumption is necessary but not sufficient to guarantee that her economic model has a pure strategy Bertrand-Nash equilibrium.

7.2.3 The Stochastic Model

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

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

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 (74).

where $\bar{\omega}_{ijt}$ are observable household and product characteristics and ϵ_{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 value error assumption appears to be made because it results in simple expressions for the firms' expectations about consumer purchase behavior found in equation (72).

The generalized extreme value error results in a nested logit model. Goldberg's choice of logit nests follows 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:

$$Pr\,(\ k \text{ buys new car } j)$$

$$\begin{aligned}
&= \ Pr(k \text{ buys a car}) \times \ Pr(k \text{ buys a new car}|\ k \text{ buys a car}) \\
&\times \ \ Pr(k \text{ buys new in segment containing } j \mid k \text{ buys a new car}) \\
&\times \ \ Pr(k \text{ buys new from } j's \text{ origin and segment} \mid k \text{ buys new in segment containing } j) \\
&\times \ \ Pr(k \text{ buys } j \mid k \text{ buys new from } j's \text{ origin and segment}).
\end{aligned} \tag{76}$$

This particular structure mirrors 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 attempts to correct 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, and avoided the IIA problem all together. Goldberg makes it clear that she prefers generalized extreme value errors because they allow her to use make maximum likelihood methods that directly deliver purchase probability estimates. Specifically, the nested logit model permits her to compute the right hand side probabilities in (76) sequentially using conventional multinomial logit software. Although Goldberg is clear that the generalized extreme value assumption is made to facilitate estimation, it is less clear how

she arrived at the particular conditioning and nesting structures that she adopts. 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 in principle 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 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.[18]

Goldberg makes two other key stochastic assumptions when she estimates her nested logit model. The first is that new car prices and non-price attributes are independent of consumers' unobserved tastes, the ϵ_{ijt}. This is a critical modeling assumption, as it is possible to imagine cases where it would not hold. Suppose, for instance, that the ϵ_{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 non-price attributes. This results in the same endogeneity problem found in neoclassical demand models. To see the parallel, imagine that ϵ_{ijt} consists of a product-time fixed effect ("quality") and noise. That is, $\epsilon_{ijt} = \xi_{jt} + \eta_{ijt}$. Since 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 firm uses when maximizing profits. Because Goldberg does not observe product quality, she would need to devise a strategy for removing it from the demand curve.

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 quality and the determinants of quality are likely to be unobserved. 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, since one would have to recognize the unknown correlation of prices and qualities when specifying the joint distribution. What Goldberg does instead is she assumes that unobserved

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

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?

A final stochastic component of the model pertains to manufacturers' marginal costs. Ignoring the issue of voluntary export restraints, the system (75) exactly identifies each product's marginal costs. Following Hausman et al (1994), she 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. It is critical to note that Goldberg's marginal costs contain several types of error. They reflect the sampling error of the CES and they contain the estimation error in the purchase probabilities. In addition, for unexplained reasons, Goldberg assumes that the costs contain an additional source of error. Specifically, she assumes that each product's marginal costs can be represented by the same linear function of product characteristics

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

where the Z_{jt} are observable product characteristics and u_{jt} are unobserved marginal costs. Goldberg does not motivate what it is about the production technology of car manufacturing that would lead to this linear conditional expectation. Following our discussion in section 4, of the difference between best linear predictors and conditional expectations, it seems that this is more a descriptive model than a structural model of costs.

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 (73) admits complicated substitution patterns among products. These substitution patterns depend, as one might hope for, 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 difficult 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, IIA is presumed 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, as other structural modelers in the same position, must extensively evaluate the economic implications of her estimates. 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 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 most other papers in this area (e.g., Porter and Hausman et al.). 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 underpredicts total new car sales, she cannot know whether this is because firms underpredicted demand or there is a problem with her specification or data.[19]

7.3 A Product-Level Demand Model

Berry (1994), Berry, Levinsohn and Pakes (BLP, 1995), and many others also have constructed differentiated product demand systems from discrete-choice models. 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, BLP (1995) only have access to product-level data. Specifically, they know a new car model's: unit sales, list price, and attributes. In all, they have twenty years of data covering 2,217 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

[19]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.

data, however, do not capture all dimensions, such as two-door versus four-door cars and standard versus luxury-equipped cars.

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:

$$q_{jt}^e = \sum_{i=1}^{M_t} Pr(\ i \text{ buys new car } \ j) = \sum_{i=1}^{M_t} Pr(P_{1t}, ..., P_{Jt}, x_{1t}, ..., x_{Jt}, \bar{\omega}_{ijt}\,;\, \theta) \qquad (77)$$

where she replaces $Pr(P, x, \bar{\omega}_{ij}; \theta)$ with her econometric model estimates of how household i's choices vary with their observable characteristics $\bar{\omega}_{ijt}$. Because Goldberg only uses household-level data, there is no guarantee that when she aggregates her probability estimates to form q_{jt}^e that they will match aggregate purchases, 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, it is as though they assume $q_{jt} = q_{jt}^e$. 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, $P(\bar{\omega}_{ijt}, \delta)$, for the household variables that they do not observe. Formally, they compute the unconditional demand functions

$$\begin{aligned} q_{jt}^e &= \textstyle\sum_{i=1}^{M_t} \int_\omega Pr(P_{1t}, ..., P_{Jt}, x_{1t}, ..., x_{Jt}, \omega; \theta)\, dP(\omega; \delta) \\ &= M_t\, \mathcal{S}_j(P_{1t}, ..., P_{Jt}, x_{1t}, ..., x_{Jt}\,;\, \theta, \delta), \end{aligned} \qquad (78)$$

where the second equality follows because 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 δ, BLP match the model's predicted expected sales $q_{jt}^e = M_t \mathcal{S}_j$ 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.

7.3.1 The Economic Model in BLP

In BLPs model, firms sell new cars directly to consumers. Firms do not price discriminate and consumers are assumed to know the prices and attributes of all new cars. There are no inter-temporal 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. The demand for the outside good is by assumption determined residually by the assumption that households buy at most one new car per year 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 $Pr(\omega; \delta)$; 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 (78) 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, since the absence of data on the outside good means that the 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}:

$$U_{ij} = \xi_j + \eta_{ij} \tag{79}$$

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

$$\mathcal{S}_j = \int_{-\infty}^{\infty} \Pi_{k \neq j} \; Pr(\, \xi_j + \tau > \xi_k) \; f(\tau) d\, \tau, \tag{80}$$

to obtain the familiar logit probabilities

$$\mathcal{S}_{jt} = \frac{\exp(\, \xi_j \,)}{\sum_{k=0}^{J} \; \exp(\, \xi_k \,)}. \tag{81}$$

In this formulation, the demand functions are

$$q_j = M \, \mathcal{S}_j(\xi_1, ..., \xi_J) \tag{82}$$

or

$$\ln q_j = \ln M + \xi_j - \ln(\sum_{k=0}^{J} \; \xi_k) \tag{83}$$

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

$$\ln q_j - \ln q_0 = \xi_j - \xi_0. \tag{84}$$

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

$$\ln q_j - \ln \left(M - \sum_{j=1}^{J} \; q_j \right) = \xi_j - \xi_0. \tag{85}$$

In general, we require at least two restrictions on the $J + 2$ unknown demand parameters $(\xi_0, \xi_1, ..., \xi_J, M)$ 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 we have to estimate M.

In their empirical work, BLP choose to fix M_t rather than restrict the ξ's or other parameters. Specifically, BLP rely on the same assumption as Goldberg and 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, and entities other than households purchase new vehicles. In principle, one could model these discrep-

ancies by assuming that the total number of US households is a noisy measure of M_t. The impact of this measurement error on the demand parameters could then be explored. To illustrate, suppose that we mistakenly use $\widetilde{M}_t = M_t + \Delta_t$ in place of M_t. Substituting $\widetilde{M}_t$ into (85) with $\xi_0 = 0$ gives

$$\ln q_j - \ln\left(\widetilde{M}_t - \sum_{j=1}^{J} q_j\right) = \widetilde{\xi}_j. \tag{86}$$

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.

Unfortunately in an actual application, we will typically not know whether we have over or underestimated M_t. This means that we will not know in which direction the bias goes on the estimated product qualities, the ξ_j's. While the availability of panel data might allow us to attempt developing a random measurement error model for M_t, in practice the nonlinearity of the demand functions in the measurement error will make it difficult in short panels to draw precise conclusions about how this measurement error impacts demand estimates. Thus, one is left with 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 two-part structure:

$$\begin{aligned} U_{ijt} &= \overbrace{\delta_{jt}}^{} + \overbrace{\omega_{ijt}}^{} \\ &= \overbrace{x_{jt}\,\beta + \xi_{jt}}^{} + \overbrace{x_{jt}\,\tilde{\nu}_i + \ln(\nu_{iy} - p_{jt}) + \eta_{ijt}}^{} \end{aligned} \tag{87}$$

The δ_{jt} includes only terms that are not household-specific. For BLP it consists of a linear function of observed (x) and unobserved (ξ) variables describing products. In this formulation, the elements of $K \times 1$ parameter vector β are interpreted as population average marginal utilities for the observed attributes, x_{jt}.

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 interaction of the household variables $\tilde{\nu}_i$ with the product attributes allows for the possibility that households do not have the same marginal utilities for attributes. While in principle one might expect that households' marginal utilities would depend on a household's income and other demographic characteristics, the lack of household data forces BLP to assume that the ν_i's are independent normal

random variables that are identically distributed in the population.[20] 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 adversion 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 the natural logarithm of a household's expenditure on the outside good, $\ln(\nu_{yi} - p_{jt})$. BLP include this term 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 they assume that the natural logarithm has a log-normal distribution. However, the lognormal distribution must be truncated differently for each model to account for the fact that the expenditure on the outside good must be positive. Put differently, their stochastic structure must have all households' incomes exceeding the price of the highest price car in their data.

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:

$$U_{i0t} = \alpha \ln \nu_{yi} + \sigma_0 \nu_{i0} + \epsilon_{i0t}.$$

Unobserved income appears here because that is the amount spent on the outside good when no new car is purchased. The parameter σ_0 is new; it represents the standard deviation of the household's unobserved preference for the outside good, ν_{i0}. To appreciate the role of ν_{i0}, recall that only differences in utilities and not utility levels are identified. In the logit model this means that the logit population shares will have the form:

$$\mathcal{S}_{jt} = \frac{\exp(\delta_{jt} + \bar{\omega}_{ijt})}{\exp(\delta_{0t} + \bar{\omega}_{i0t}) + \sum_{k=j}^{J} \exp(\delta_k + \bar{\omega}_{ikt})}. \tag{88}$$

which is equivalent to (assuming without loss of generality $\beta = 0$)

$$\mathcal{S}_{jt} = \frac{\exp((\xi_{jt} - \sigma_0\nu_{0i}) + (\bar{\omega}_{ijt} - \alpha \ln \nu_{yi}))}{1 + \sum_{k=j}^{J} \exp((\xi_{kt} - \sigma_0\nu_{0i}) + (\bar{\omega}_{ikt} - \alpha \ln \nu_{yi}))}. \tag{89}$$

This expression reveals that including an idiosyncratic taste for the outside good is equivalent to a model in which we included a household-specific constant in the unobserved ξ_{jt}. Thus, if the unobserved attribute were product quality, the ν_{0i} would

[20] BLP and students have in follow on papers explored alternatives to this structure. Using micro data, as in Goldberg (1995), BLP (1998) allow consumers' marginal utilities to depend on observable and unobservable household attributes.

linearly translate each household's quality ratings. Households therefore would all agree on the difference in any two cars' qualities, but households would disagree on the overall quality of new cars. In particular, 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. Thus, holding the population average qualities of new cars fixed, increases in the standard deviation of ν_{i0} have the effect of forcing substitution away from all new car models toward the outside good.

7.3.3 More on the Econometric Assumptions

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

7.3.4 Functional Form Assumptions for Price

A first critical component of the specification of any choice model is the assumption made about how product prices affect 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:

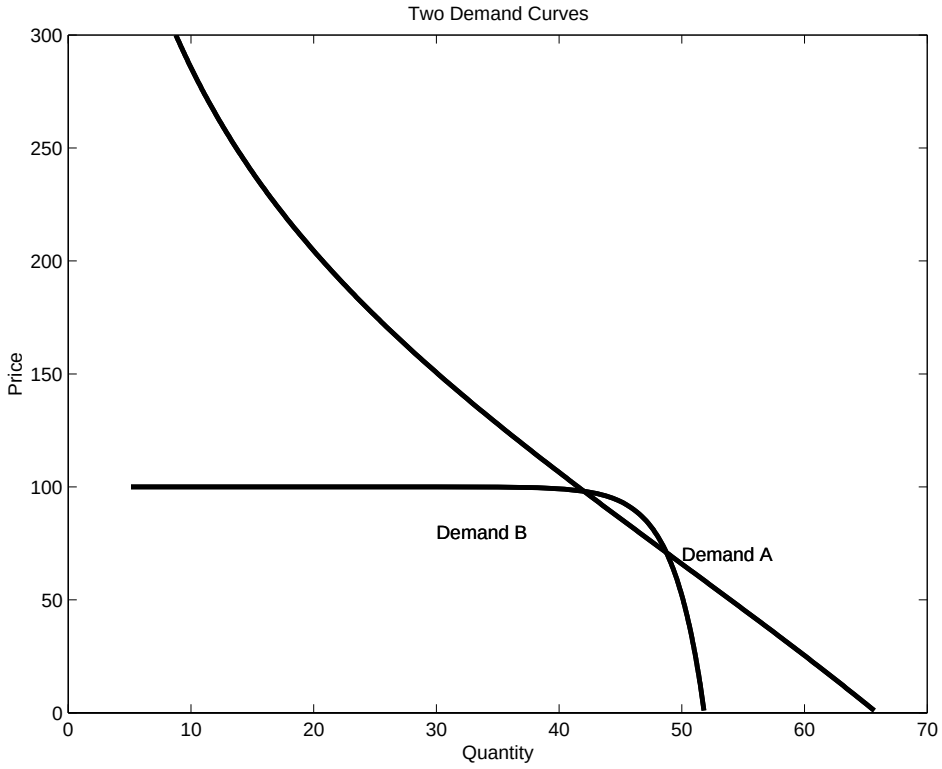
$$\ln q_{jt} = \ln M_t + g(p_{jt}) + \tilde{\delta}_{jt} - \ln\left(1 + \sum_{k=1}^{J} \exp(\, g(p_{kt}) + \tilde{\delta}_{kt}\right) \tag{90}$$

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

$$\frac{\partial \ln q_{jt}}{\partial \ln p_{kt}} = \begin{cases} \frac{\partial g(\delta_j t)}{\partial p_{jt}}\, p_{jt}\,(1 - \mathcal{S}_{jt}) & k = j \\ -\frac{\partial g(\delta_{jt})}{\partial p_{kt}}\, p_{kt}\, \mathcal{S}_{kt} & k \neq j \end{cases} \tag{91}$$

These expressions reveal how the extreme value error assumption and the choice of $g(\cdot)$ can dramatically impact the structure of the own-price and cross-price elasticities that enter the price-markup equations. If price enters logarithmically (e.g., $g(p_{jt}) = \alpha \ln p_{jt}$), then the own- 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 for finite attributes there always be someone who will buy a make – 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 (2001), 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 if the researcher is making predictions or performing counterfactual calculations for prices outside this range of prices, the difference between the two demand curves can 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.



7.3.5 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 (LARGE=1);
- utility for the large and small cars have the form

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

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

This two-parameter utility specification perfectly explains the large car out-selling the small car 3-to-1 by assigning a positive marginal utility to size of the car ($\beta_0 = -3.871$ and $\beta_1 = 1.098$).[21] Although the mean utility specification predicts consumers prefer larger to smaller cars, the infinite support of the extreme value error ϵ_{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_1 for size and those with a taste β_2. Because we can potentially explain the three market shares with just two parameters, set $\beta_0 = 0$. 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. When we fit the predicted shares to the actual shares in this case, however, we find that 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 to size ($\beta_1 = 3.9836$). Thus, when consumers' marginal utilities are unconstrained, the choice model may explain the purchase of an inferior product by estimating that some consumers have negative marginal utilities for otherwise attractive attributes.

[21] Explain two parameters.

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. Thus, at the same price, Peter and Frank may both want faster computers, but speed is much more important to Frank. In the utility specifications above, we can think of a pure vertical model as one in which consumers have positive, but not necessarily the same marginal utility for the size of a car. In horizontal differentiation models, consumers differ in their opinion about an attribute, and thus rank products differently. For example, at the same price, Peter and Frank would have different views on the desirability of high cholesterol snacks.

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. For instance, Peter and Frank might both want fast computers loaded with memory. They may choose different computers because Peter cares more intensely about memory while Frank cares more intensely about speed. Alternatively, differences in rankings also can be explained by some consumers having negative marginal utilities for some attributes. Peter and Frank again might both want fast computers loaded with memory, but computers also come with different operating capabilities (e.g., Macintosh OS versus Windows) for which Peter and Frank might have opposing preferences.

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 econometric results, the mean marginal utility for miles per dollar is -.122 with an estimated standard deviation of 1.050. Thus, roughly 54 percent of consumers "dislike" this fuel economy attribute. While it is debatable whether this estimate is high or low, the critical observation is that to explain the aggregate sales data, their model ascribes considerable heterogeneity to consumers' tastes.

Since inferences about consumer heterogeneity are predicated on maintained functional form assumptions, it seems imperative that some effort should go into explor-

ing 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 (2000), Berry (2001), Bajari and Benkard (2001a and 2001b) 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.

To see what is involved computationally, recall that the unconditional purchase probabilities have the form:

$$\mathcal{S}_{jt} = \int_{\nu} \frac{\exp(\,\delta_{jt} + \omega_j(\nu)\,)}{1 + \sum_{k=1}^{J} \exp(\,\delta_{kt} + \omega_k(\nu)\,)}\, \phi(\nu, \delta)\, d\nu \tag{92}$$

In BLP's case, $\phi(\cdot)$ is a product of standard normal densities and δ is a parameter vector containing the standard deviations of the unobserved marginal utilities.[22] As there is no analytic simplification for this integral, BLP use simulation methods to evaluate it. Their paper contains a lengthy treatment of the simulation methods use to evaluate this integral, and the significance of simulation error for their estimates. The fact that one cannot analytically evaluate consumers' choice probabilities is not a problem in and of itself. It does, however, make it more difficult for the researcher to evaluate and report to other researchers just how flexibile the choice model is over and above a conventional logit model.

7.3.6 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

$$\ln q_j - \ln\left(M - \sum_{j=1}^{J} q_j\right) = \xi_j. \tag{93}$$

From this equation we see that the ξ_{jt} act as demand "errors" that insure that the econometric choice model's predicted market shares match the observed market shares. In BLP's model it is essential that the predicted and observed market shares match. This is because BLP's theoretical model presumes that (uncondition-

[22] This "mixed normal-logit model" has a prior history. See McFadden and Train (1998) for an overview of early work.

ally) each consumer's decision can be represented by the same multinomial choice probabilities: $(\mathcal{S}_0, \mathcal{S}_1, ..., \mathcal{S}_J)$. Thus, with a sample size, M, of approximately 100 million, there should be no 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 (85) become

$$\ln q_j - \ln\left(M - \sum_{j=1}^{J} q_j\right) = x_j\beta + \xi_j. \tag{94}$$

Assuming M is known, we have J linear equations in $J+K$ unknowns: $(\xi_1, ..., \xi_J, \beta)$. 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.[23] 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: $E(\xi_j|x_1, ..., x_J) = 0$. This moment condition is useful because it suggests least squares can be used to estimate the marginal utilities in (94). With least squares, the K population moment conditions $E(\xi_j|x_j) = 0$ are replaced by the K sample moment conditions $\sum_{t=1}^{J} x_t\xi_j = 0$. These K sample moments result in the estimated product qualities having $J-K$ degrees of freedom.

While imposing the population moment condition $E(\xi_j|x_1, ..., x_J) = 0$ 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, 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 condi-

[23] 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.

tioning variables $x_1, ..., x_J$. Of course, this means that they must introduce another 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.[24] Indeed, given the $J \times K$ attributes in $x_1, ..., x_J$, BLP BLP have many choices for instruments, making their model highly over-identified. Because of concerns with near collinearity (which recall were also are present in neoclassical demand systems), BLP limit the number of instruments they use in forming their moment conditions for the ξ's. In particular, they use: all non-price attributes of the car; the sum of all non-price attributes of cars made by the same manufacturer; and the sum of all non-price attributes of cars made by competing manufacturers. As BLP have five non-price attributes, these choices provide 15 moment conditions.

One question that BLP do not explicitly address is why it is acceptable to use non-price attributes to construct instruments. This is the same issue that arose in our discussion of neoclassical demand systems. One might well imagine here 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 would require instruments as well.

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 (92), in general it is not possible to compute closed form expressions for the ξ_{jt} that enter the population moment conditions. This means in practice that the researcher must numerically invert equation (80) to solve for the ξ_{jt}. While the integral in (92) 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 (92), which they then invert to obtain the ξ_{jt} for candidate values of the taste parameters.

7.3.7 The Cost Specifications

To this point, we have said little about the cost side. In principle, one could estimate the demand parameters without using information about the supply side. BLP appear to add the supply side for two reasons. First, it contributes variables that can be used in the orthogonality conditions that identify the demand parameters. Second, it allows

[24] For example in the cross section logit model we can replace the moment condition $E(\xi_j|p_j) = 0$ with $E(\xi_j|x_{k1}) = 0$, 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.

them to estimate markups for different models and calculate the profit potential of attributes, just as Hausman and Goldberg do. The latter benefit seems much more important than the former since the cost side contributes just two additional variables to the model (a Time Trend and Miles per gallon). Following the approach discussed above for constructing demand error instruments, BLP potentially can add 21 (seven instruments times 3) sample moment conditions for the cost-side error. Because of collinearity concerns, they drop two of these moment conditions in estimation.[25]

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.

7.4 Summary

BLP report estimates for several demand models. They also 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. It is much more difficult for the authors to provide a sense for 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.[26] Ongoing work by others is exploring the consequences of other assumptions in these models, and we leave it to others to survey this work.[27]

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,

[25] 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.

[26] For example, Berry(2001) and Berry, Levinsohn and Pakes (1998).

[27] For example, Ackerberg and Rysman (2000), Bajari and Benkard (2001a), and Bajari and Benkard (2001b).

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. This is true both on economic and statistical grounds. 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 grounding in representative agent formulations. There also to date there have been few attempts made to investigate the finite sample or asymptotic performance of different estimation procedures.[28] Despite these open issues, academics and antitrust authorities currently are using these models to perform welfare calculations, and evaluate mergers and other antitrust issues.

8 Models with Private Information: Auctions

[Section not yet in this draft.]

9 Econometric Models of Entry, Exit and the Number of Firms in a Market

IO economists have long been fascinated with measuring industrial concentration (e.g., NBER 1955). This work has played critical roles in the development of US trade policy, tax policy and, in particular, antitrust and merger policy. For example, recent US antitrust and merger guidelines have partly assessed the competitiveness of markets using market concentration measures.

Debates in the 1970s and 1980s about the usefulness of market concentration measures led IO economists to think more deeply about connections between competition and market concentration. Two related theoretical literatures emerged from these explorations. One literature addresses the question: "How many firms must be in a market to have effective competition?" The second asks: "What factors encourage or deter firms to enter and exit markets?" Both literatures emphasize that in oligopoly markets, the answers to these questions often hinge delicately on what one assumes about firms' strategic interactions. Indeed, much of the recent theoretical literature

[28] 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 (2002).

in this area has tended to emphasize that it is strategic behavior, rather than rather than technological or other factors, that has the greatest impact on market structure.

Unfortunately, very little empirical work exists that attempts to separate out the importance of technological, demand and strategic factors for firms' incentives to enter markets and compete. In large part this is because to do so, one requires a much more complicated oligopoly model than those considered in Sections 5, 6 and 7. Only recently have empirical researchers begun to make progress in developing structural econometric models that can speak to strategic models of entry and entry deterrence. In this subsection we outline some of the econometric issues that arise in modeling oligopolistic markets where strategic behavior may be a factor.[29] The main point of this section is to show that the potential for strategic behavior introduces complex structural modeling issues. Several of these issues are generic, such as the possible multiplicity of equilibrium market structure outcomes, and these issues have only begun to be addressed by the literature.

Before exploring a specific structural model of market concentration, it is useful to have a broader sense of the economic issues that models of market concentration and competition might address.

9.1 An Example

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.[30] One does not have to look far to understand why. Even though travelers often encounter wide disparities in fares (per seat miles) over time, across routes and even for seats on the same flight, fares do not vary as much across competing carriers. Industry critics contend that such patterns are obvious evidence of ineffective competition. They also argue that the 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 normal 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 to have high (per seat mile)

[29] See also Berry and Reiss (2002).

[30] See for example Borenstein (1992), Brueckner (1992), Morrison (1996), Ott (1990), and Windle (1993).

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 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 can't generally. The main reason is that much of what the theory points us toward is unobservable. We (the 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, 6 and 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, 6 and 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. The theory says 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 post 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 among 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 may have about post-entry competition.

9.2 The Economic Model

Our airline example makes three points that are worth re-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, it is important for us to have sufficient data on a cross section of city-pair markets where firms have similar costs and face similar demands. One also could imagine as an alternative assembling time series data on the same market over time. Such data would have the advantage of holding constant market-specific conditions. The use of time series data, however, raises new modeling issues. Now researchers must worry about whether firms' decisions are 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, and as such they tend to use cross-section data on the number of firms in different, related markets. That is, most market structure models explain the number of firms in a market rather than firm turnover.

A typical static model follows the approach of the competition models discussed in Sections 5, 6 and 7. The researcher builds a model of competition from primitive demand, cost and behavioral assumptions. Unlike the models in prior sections, assumptions about fixed costs now play an important role in these models, as they help determine which set of firms will produce, and therefore which set of marginal conditions to analyze. 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 (1991, 1991b) to highlight some of these issues and the strengths and weaknesses of structural models.

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; they do not have price or quantity information. Thus, absent a structural model, the best they can do is summarize how the joint distribution of entry and exit rates vary with industry characteristics. (See for example Dunne, Roberts and Samuelson (1988).) In adopting a structural approach, Bresnahan and Reiss' modeling is complicated by the fact that entry and exit are discrete events. Thus, their structural models will not typically involve marginal conditions, such as is done in the models discussed in Sections 5, 6 and 7. Instead, they must base their inferences solely 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. This payoffs to the players of the game are:

	Stay Out (0)		Enter (1)	
Stay Out (0)	$\Pi_1(0,0)$	$\Pi_2(0,0)$	$\Pi_1(0,1)$	$\Pi_2(0,1)$
Enter (1)	$\Pi_1(1,0)$	$\Pi_2(1,0)$	$\Pi_1(1,1)$	$\Pi_2(1,1)$

where the $\Pi_k(i,j)$ represent the profits firm k earns when firm 1 plays i and firm 2 plays j (a zero denotes the action "Stay Out" and a one denotes "Enter"). In most 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., firm 1 enters and firm 2 stays out) but does not observe the firms' economic profits (the $\Pi_k(i,j)$). 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, 6 and 7, that Bresnahan and Reiss will have to introduce considerable structure in order to draw inferences about firm profits and behavior from discrete outcomes.

9.3 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 $a_1, a_2, ..., a_N$ of the $N \geq n$ potential entrants to draw inferences about firms' profit functions, i.e.,

$$\Pi_k(a_1, a_2, ..., a_N, Z, \theta). \tag{95}$$

Here Z 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 (95). 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

$$\Pi_k(a_1, a_2, ..., a_N, Z, \theta) = b \left(\frac{S}{\sum_{j=1}^{N} a_j + 1} \right)^2 - F \tag{96}$$

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 $\Pi_i \geq 0$. 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

$$\frac{S^2}{(n+1)^2} \geq \frac{F}{b} \geq \frac{S^2}{(n+2)^2}.$$

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 $S_1, S_2, ...$ 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

$$\frac{S_2^2}{S_1^2} = \frac{9}{4} = 2.25 \tag{97}$$

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 (1991)).

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 need to observe to draw sharper inferences.

9.4 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 N, 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 consist of a variable profit and a fixed cost term

$$\bar{\Pi}_k(a, Z, \theta) = VP_i(\, a, Z, \theta) - F_i(a, Z, \theta). \tag{98}$$

Here, a is a vector describing the N potential entrants' entry actions, VP denotes variable profits and F fixed costs. 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 (98). 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 (98), 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

$$VP_i(\, a, Z, \theta) - \tilde{F}_i(a, Z, \theta) \geq \epsilon_i(a) \tag{99}$$

where a tilde ($\tilde{}$) above a variable denotes the economic quantity 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 the 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

$$\begin{array}{rcl} \Pi_1(a_1^*, a_2^*) & \geq & \Pi_1(a_1, a_2^*) \\ \text{and} & & \\ \Pi_2(a_1^*, a_2^*) & \geq & \Pi_2(a_1^*, a_2) \end{array} \tag{100}$$

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 (0)	Enter (1)
Stay Out (0)	$\Pi_1(0,0) \quad \Pi_2(0,0)$	$\Pi_1(0,1) \quad \Pi_2(0,0) + \Delta_0^2$
Enter (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

$$\begin{array}{rcl} a_1 = 0 & \iff & \Delta_0^1 + a_2\, \Delta_1^1 \leq 0 \\ a_2 = 0 & \iff & \Delta_0^2 + a_1\, \Delta_1^2 \leq 0 \end{array} \tag{101}$$

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

$$\begin{array}{rcl} \Delta_0^i & = & \frac{(\alpha - c)^2}{4\, b} - F = g(\alpha, c) - F \\ \Delta_1^i & = & \frac{5(\alpha - c)^2}{36\, b} = h(\alpha, c) \end{array} \tag{102}$$

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

$$a_i = \begin{cases} 1 & \text{if } y_i^* = g(\alpha, c) - F + a_j\, h(\alpha, c) - \epsilon_i \geq 0 \\ 0 & \text{if } y_i^* = g(\alpha, c) - F + a_j\, h(\alpha, c) - \epsilon_i < 0 \end{cases} \tag{103}$$

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

$$a_i = \begin{cases} 1 & \text{if } y_i^* = X\,\beta_i + a_j\,\delta - \epsilon_i \geq 0 \\ 0 & \text{if } y_i^* = X\,\beta_i - F + a_j\delta - \epsilon_i < 0 \end{cases} \tag{104}$$

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 non-uniqueness problem, which is to aggregate the non-unique outcomes (in this case the non-unique 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.

9.5 Estimation

Turning now to estimation, Bresnahan and Reiss (1991) 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 each of a number of 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

$$\Pi_i(1, a_2) = (g(Z, \beta) + a_2\, h(Z, \delta))\, S - F(a_2) - \epsilon \tag{105}$$

From this relation, Bresnahan and Reiss identify entry thresholds for a monopolist and a duopoly. That is, the entry thresholds equal

$$S(a_2) = \frac{F(a_2) - \epsilon}{g(Z, \beta) + a_2\, h(Z, \delta)} \tag{106}$$

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:

$$\frac{S(1)}{S(0)} = \frac{F(1)}{F(0)}\, \frac{g(Z, \beta)}{g(Z, \beta) + h(Z, \delta)} \tag{107}$$

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 (98) and then forming the ratio (107). 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.

9.6 Epilogue

A number of researchers have extended Bresnahan and Reiss' models and explored alternatives. 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 reduced form probit models and the like are inadequate for modeling entry and exit decisions.

10 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 as a field 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 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 descriptive and structural models. We should emphasize once more that we see great value in both of these enterprises. 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, for example, 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. and M. Rysman (2000). "Unobserved Product Differentiation in Discrete Choice Models: Estimating Price Elasticities and Welfare Effects". Boston University Department of Economics Working Paper.

Amemiya, T. (1974), "Multivariate Regression and Simultaneous Equation Models When the Dependent Variables are Truncated Normal," Econometrica, 42, 999-1012.

Applebaum, E. (1982), "Estimation of the Degree of Oligopoly Power," Journal of Econometrics, 19, 287-299.

Bain, J. S. (1956), Barriers to New Competition, Their Character and Consequences in Manufacturing Industries. Harvard University Press, Cambridge.

Baker, J. and Bresnahan, T. (1988), "Estimating the Demand Curve Facing a Single Firm", International Journal of Industrial Organization, 6, 283-300.

Bajari, P. and L. Benkard. (2001a), "Discrete Choice Models as Structural Models of Demand: Some Economic Implications of Common Approaches". Stanford Graduate School of Business Working Paper.

Bajari, P. and L. Benkard. (2001b), "Demand Estimation with Heterogeneous Consumers and Unobserved product Characteristics: A Hedonic Approach". Stanford Graduate School of Business Working Paper.

Becker, G. (1962), "Irrational Behavior and Economic Theory," Journal of Political Economy, 70, 1-13.

Berry, S. (1994), "Estimating Discrete-Choice Models of Product Differentiation," RAND Journal of Economics, 25(2), 242-262.

Berry, S. (2001), "Estimating the Pure Hedonic Choice Model", Yale Department of Economics Working Paper.

Berry, S., J. Levinsohn and A. Pakes (1995), "Automobile Prices in Market Equilibrium," Econometrica, 63(4), 841-890.

Berry, S., J. Levinsohn and A. Pakes (1998), "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," NBER Working Paper # 6481.

Berry, S., O. Linton and A. Pakes (2002), "Limit Theorems for Estimating the Parameters of Differentiated Products Demand Systems". Yale Department of Economics Working Paper.

Berry, S., P. Reiss (2002), "Empirical Models of Entry and Exit". In process.

Blackorby, C., D. Primont and R. Russell (1978), *Duality, Separability and Functional Structure.* North Holland, Amsterdam.

Borenstein, S. (1992), "The Evolution of U.S. Airline Competition," Journal of Economic Perspectives.

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

Bresnahan, T. (1982), "The Oligopoly Solution Concept is Identified," *Economics Letters*, 10, 87-92.

Bresnahan, T. (1987), "Competition and Collusion in the American Automobile Market: The 1955 Price War," Journal of Industrial Economics, 45(4, June) (special issue), 457-482.

Bresnahan, T. (1996), "Comment", in T. Bresnahan and R. Gordon, *The Economics of New Goods.* NBER Studies in Income and Wealth Volume 58. The University of Chicago Press, Chicago.

Bresnahan, T. (1997), "The Apple-Cinnamon Cheerios War: Valuing New Goods, Identifying Market Power, and Economic Measurement." Stanford University mimeograph.

Bresnahan, T. (1989), "Empirical Studies of Industries with Market Power", R. Schmalensee and R. Willig, eds., The Handbook of Industrial Organization, North Holland, Amsterdam.

Bresnahan, T. and P.C. Reiss (1985), "Manufacturer and Dealer Margins," Rand Journal of Economics, , 309-323.

Bresnahan, T. and P.C. Reiss (1991), "Entry and Competition in Concentrated Markets," Journal of Political Economy, 99(5), 977-1009.

Bresnahan, T. and P.C. Reiss (1991b), "Empirical Models of Discrete Games," Journal of Econometrics, 48, 57-81.

Brown, A. and A. Deaton (1972), "Models of Consumer Behavior: A Survey," Economic Journal, 82, 1145-1236.

Brueckner, J.K. (1992), "Fare Determination in Hub and Spoke Networks," Rand Journal of Economics, 23 #3, 309-323.

Burns, A.F. and W.C. Mitchell (1946), *Measuring Business Cycles*, NBER, New York.

Camerer, Colin (1995), "Individual Decision Making," Chapter 8 in J. Kagel and A. Roth, eds., The Handbook of Experimental Economics, Princeton University Press, Princeton.

Caves, R., Whinston, M., Hurwitz, M. (1991), "Patent Expiration, Entry, and Competition in the U.S. Pharmaceutical Industry," Brookings Papers on Economic Activity: Microeconomics, 1-48.

Comanor, W.S. and Wilson, T.A. (1967), "Advertising, Market Structure and Performance," Review of Economics and Statistics, 49, 423-40.

Corts, K. (1999), "Conduct Parameters and the Measurement of Market Power, *Journal of Econometrics*, 88 (2), 227-250.

Davis, P. (2000), "Demand Models for Market-Level Data", MIT Sloan Working Paper.

Deaton, A. and J. Muelbauer (1980), *Economic and Consumer Behavior*, Cambridge: Cambridge University Press.

Demsetz, H. (1973), "Industry Structure, Market Rivalry and Public Policy," Journal of Law and Economics, 16, 1-9.

Dixit, A. and J. Stiglitz, (1977), "Monopolistic Competition and Optimum Product Diversity", *American Economic Review*, 67, 297-308.

Dunne, Roberts & Samuelson (1988), "Patterns of Firm Entry and Exit in U. S. Manufacturing Industries," RAND Journal of Economics, 19(4, Winter), 495-515.

Engel, E. (1857), "Die Productions-und Consumptionsverhltnisse des Knigreichs Schsen," in Zeitschrift des Statischen Bureaus des Kniglich S chsischen Ministerium des Inneren, Nos. 8 and 9.

Goldberg, Penny K., (1995), "Product Differentiation & Oligopoly in International Markets: The Case of the U.S. Automobile Industry," Econometrica, 63(4), 891-952.

Goldberger, Art (1991), *A Course in Econometrics*, Harvard University Press, Cambridge, MA.

Gollop, F. and Roberts, M. (1979), "Firm Interdependence in Oligopolistic Markets," Journal of Econometrics, 10, pp. 313-331.

Gorman, W. (1959), "Separable Utility and Aggregation", Econometrica, 27, 469-481.

Gorman, W. (1970), "Two-Stage Budgeting", in C. Blackorby and A. Shorrocks, eds., *Separability and Aggregation. Collected Works of W.M. Gorman. Volume I.* Clarendon Press, Oxford.

Green, J. and R. Porter (1984), "Noncooperative Collusion Under Imperfect Price Information," Econometrica, 52, 87-100.

Griliches, Z. (1957), "Hybrid Corn: An Exploration into the Economics of Technical Change," Econometrica, 25, 331-346.

Griliches, Z. (1986), "Economic Data Issues," Chapter 25 in Z. Griliches and M.D. Intriligator eds., Handbook of Econometrics, Vol 3., North Holland, Amsterdam.

Haile, P. (2001), "Auctions with Resale Markets: An Application to U.S. Forest Service Timber Sales," American Economic Review, 91, 399-427.

Haavelmo, T. (1944), "The Probability Approach in Economics," Econometrica, iii-vi, and 1-115.

Hanemann, W. (1984), "Discrete/Continuous Models of Consumer Demand", Econometrica, 52, 541-561.

Hausman, J., (1996), "Valuation of New Goods under Perfect and Imperfect Competition," in T. Bresnahan and R. Gordon, *The Economics of New Goods.* NBER Studies in Income and Wealth Volume 58. The University of Chicago Press, Chicago.

Hausman, J., G. Leonard, and D. Zona. (1994), "Competitive Analysis with Differentiated Products", Annales dEconometrie et de Statistique, 34, 159-180.

Heckman, J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System," Econometrica, 46, 931-959.

Hendricks & Porter (1988), "An Empirical Study of an Auction with Asymmetric Information," American Economic Review, 78(5, December), 865-883.

Hood, W.C. and T.C. Koopmans (1953), Studies in Econometric Method, Cowles Commission Monograph No. 14, John Wiley & Sons, New York.

Laffont, Ossard, and Vuong, (1995), "Econometrics of First-Price Auctions," Econometrica, 63(4), 953- 980.

Lau, L. (1982), "On Identifying the Degree of Industry Competitiveness From Industry Price and Output Data," *Economics Letters*, 10, 93-99.

Lerner, A. (1934), "The Concept of Monopoly and the Measurement of Monopoly Power," *Review of Economic Studies*, 11, 157-175.

Lindh, T. (1992), "The Inconsistency of Consistent Conjectures," Journal of Economic Behavior and Organization, 18, 69-90.

Markham, J.W. (1952), Competition in the Rayon Industry, Harvard University Press, Cambridge.

Martin, Stephen (1993), *Advanced Industrial Economics*, Blackwell, Oxford.

Mitchell, W.C. (1926), Business Cycles: The Problem and Its Setting, NBER, New York.

Morrison, S.A. (1996), "Causes and Consequences of Airline Fare Wars," Brookings Papers on Economic Activity: Microeconomics, 85-124.

National Bureau of Economic Research, Business Concentration and Price Policy, Princeton University Press, Princeton, NJ.

Nevo, A. (1997), "Mergers with Differentiated Products: The Case of Ready-to-Eat Cereal Industry", *Rand Journal of Economics*, 31, 395-421.

Ott, J. (1990), "Justice Dept. Investigates Carriers' Pricing Policies," Aviation Week and Space Technology, 133 #3, 18-20.

Paarsch, H. (1992) "Deciding Between the Common and Private Values Paradigms in Empirical Models of Auctions," Journal of Econometrics, 51(1/2), 191-216.

Paarsch, H. (1997), "Deriving an Estimate of the Optimal Reserve Price: An Application to British Columbia Timber Sales," Journal of Econometrics, 78, 333-357.

Pinske, J., M. Slade and C. Brett (2000), "Spatial Price Competition," University of British Columbia Department of Economics Working Paper; forthcoming in Econometrica.

Pinske, J., and M. Slade (2001), "Mergers, Brand Competition, and the Price of a Pint", University of British Columbia Department of Economics Working Paper.

Polak, R.A., and Wales, T.J. (1992), *Demand System Specification and Estimation*, Oxford University Press, New York.

Porter, R.. (1983), "A Study of Cartel Stability: The Joint Executive Committee, 1880-1886," Bell Journal of Economics, 14(2, Autumn), 301-314.

Reiss, P.C. and P. Spiller (1989), "Competition & Entry in Small Airline Markets," Journal of Law & Economics, 32(2, October, part 2), S179-S202.

Riordan, M. (1985), "Imperfect Information and Dynamic Conjectural Variations," *Rand Journal of Economics*, 16, 41-50.

Rosse, J. (1970), Estimating Cost Function Parameters Without Using Cost Data: Illustrated Methodology," Econometrica, 38, 256-275.

Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Sims, C. A. (1980), "Macroeconomics and Reality," Econometrica, 48, 1-47.

Spiller, P. and Favaro, E. (1984), "The Effects of Entry Regulation on Oligopolistic Interaction: The Uruguayan Banking Sector," Rand Journal of Economics, 15, 244-254.

Stigler, G. (1964), "A Theory of Oligopoly", Journal of Political Economy, 93, 44-61.

Trajtenberg, M. (1989), "The Welfare Analysis of Product Innovations, with an Application to CAT Scanners," Journal of Political Economy, 97(2), 444-479.

Ulen, T.S. (1978) Cartels and Regulation. Unpublished Ph.D. dissertation, Stanford University.

Waterson, M. (1984) Economic Theory of the Industry, Cambridge University Press: Cambridge.

Windle, R. (1993), "Competition at 'Duopoly' Airline Hubs in the U.S.," Transportation Journal, 33 #2, 22-30.

Wolak, F. (1994), "An Econometric Analysis of the Asymmetric Information Regulator-Utility Interaction, Annales D'Economie et de Statistique, 34, 12-69.

Wolak, F. (1997), "The Welfare Impacts of Competitive Telecommunications Supply: A Household-Level Analysis," Brookings Papers on Economic Activity: Microeconomics,269-340.