Introduction to Structural Estimation in IO*

Luca Maini

June 30, 2026

Contents

*These notes were prepared for a 2-day mini course offered to early-stage graduate students in preparation for taking a graduate IO course. Please email maini@hcp.med.harvard.edu for corrections.

Overview

The main goal of this course is to provide a basic introduction to Industrial Organization (IO) and structural estimation. As such, it can be viewed as preparation for a second-year graduate IO course centered around these methods. Since this course is designed for students in a Health Policy PhD program, I will mostly rely on examples from health economics. The course covers three main topic-areas:

1. A basic overview of theoretical IO models and concepts
2. An introduction to structural estimation, why and how it differs from reduced-form approaches, and a discuss of its advantages and disadvantages.
3. A discussion of what "structural identification" means in practice, and how to read and interpret the results of structural papers.

In the last two sections, I also provide an informal discussion of the most common estimators used in structural estimation (Minimum Distance Estimators—or MDE—and Maximum Likelihood Estimators—or MLE).

The course is generally organized as two lectures. The first lecture covers Sections 1 and 2. The second lecture covers Sections 3 and 4.

1 Introduction to Industrial Organization

1.1 History

The goal of this section is to present the current conventions of research in industrial organization (IO), and to motivate why and how they arose historically. IO, as a field, studies how market structure affects equilibrium outcomes. For example, how much higher are prices under a monopoly regime than under a duopoly? How big is the markup from product differentiation? How do decreasing marginal costs affect firm size and equilibrium prices? Is innovation higher under monopoly or under perfect competition?

Roughly speaking, research on these topics consists of three waves (and maybe a fourth one):

1. Early empirical wave, or the **"structure-conduct-performance"** paradigm (SCPP). This is the idea that **market structure** (e.g., number of sellers, mode of competition, degree of product differentiation, cost structure, etc.) affects **conduct** (e.g., strategic firm choices like prices, quantities, advertising, R&D, etc.). Early research regressed "conduct" variables on "structure" variables, often across industries (for an overview, see Schmalensee, 1989).

2. **Theoretical wave**: starting in the 1970s, this wave consists of mainly theoretical papers that analyze oligopoly markets. Its main contribution is in the treatment of more complex concepts, such as dynamics and asymmetric information. The models and results developed during this time form the backbone of advanced undergraduate and graduate courses on theoretical IO. The emergence of this work can trace its roots in the dissatisfaction with the rudimentary empirical approach to analyzing markets, and the desire to provide more accurate descriptions of firm strategic decision-making. The work was theoretical almost by necessity, due to lack of data and the computing power necessary to analyze it. The best resource to learn about these models is Jean Tirole's textbook (Tirole, 1988).

3. The **New Empirical Industrial Organization (NEIO)**: the same critiques moved by the theoretical wave also spurred researchers to think about different ways of conducting empirical work. For example, Rosse (1970) backs out marginal costs for newspapers from a profit-maximization assumption and a demand model. But, by and large, the computing power required to do this type of analysis did not exist, and neither did the data (Rosse hand-collected prices and quantities from industry publications and by cold-calling newspapers). So this work remained sparse until the 1980s, when computers caught up. Early papers in this wave include Bresnahan (1981); Green and Porter (1984); Pakes (1986); Suslow (1986); Rust (1987). The NEIO critiqued certain tenets of the early SCPP empirical approach: the idea that costs are easily observed in data, and the notion that differences across industries can be boiled down to a handful of observed parameters. Tim Bresnahan's chapter in the first handbook of IO provides a good overview of the motivations behind this wave (Bresnahan, 1989).

4. A new wave? Market-wide empirical work. The main weakness of NEIO is that it almost

exclusively focuses on single industries. This approach, which was chosen to allow for customization of modeling to the specific settings of an industry (and is still by far the prevalent paradigm in the field), nonetheless struggles to answer questions about the overall market. Recently, such questions have popped up with more prominence: have markups risen over time? If so, why? Do cross-market mergers matter for antitrust purposes? Under what conditions? These are questions that can only be answered looking at the economy as a whole (though case-studies can be quite helpful at generating hypotheses). Some empirical papers have leveraged newly available data and computing power to make some progress in this space (see, e.g., Pellegrino, 2025)

Current empirical research in IO is still largely based on the tenets that shaped the NEIO revolution:

- Attention to industry-specific structure and setting in modeling.
- Acute awareness of unobservable data, especially cost.

In addition, as empirical work has gotten progressively more sophisticated, the field is taking an increasingly skeptical view of certain assumptions, especially as they pertain to the statistical properties of the model parameters (e.g. the distribution of error terms) made for the sake of simplifying estimation. While these assumptions were common early on in the literature, they are now viewed unfavorably in many cases.

1.2 Theoretical Concepts and Models in Industrial Organization

The basic model that underpins most of Industrial Organization is a model of competition between multiple firms. You have almost certainly seen at least one of these models, the Cournot model of competition over quantity.

In the simplest version of the model, N identical (i.e., with the same cost function) firms compete by simultaneously choosing how many units to produce of a given homogeneous good. Once all firms have made their choices, prices adjust so that the market clears. This translates to the following maximization problem:

$$\max_{q_j} q_j \times p\left(q_j + q_{-j}\right) - c\left(q_j\right) \tag{1}$$

where $q_{-j} = \sum_{k \neq j} q_k$ is the quantity produced by other firms, $p\left(q_j + q_{-j}\right)$ is the market-clearing price, and $c\left(q_j\right)$ is the cost of producing q_j units.

We solve this model by taking a FOC:

$$\left[q_j\right] : p\left(q_j + q_{-j}\right) + \frac{\partial p}{\partial q_j} \times q_j - \frac{\partial c}{\partial q_j} = 0 \tag{2}$$

The solution comes from combining the FOCs for each j and solving them jointly.

All models in IO are essentially variations on this theme:

- What if costs are heterogeneous?
- What if firms set price rather than quantity?
- What if firms have to pay a fixed cost to enter the market?
- What if products are differentiated?
- What if firms can price-discriminate?
- What if one firm has more information than the others about the market?
- What if firms have to buy inputs of production from other firms?
- What if firms can merge?
- What if firms compete over multiple periods?
- What if cost functions decline the higher the quantity produced (e.g. learning-from-doing?)
- What if demand today affects demand tomorrow (e.g., inertia/learning)?
- What if firms can invest to innovate?
- What if the market is regulated by the government?

And so on and so forth.

While these models can become very complicated, and have vastly different implications from one another, they all share the same two conceptual goals. First, to show that firms are strategic actors that will react to the competitive environment they face. Second, that the characteristics of said environment matter greatly for the final outcome.

1.3 From theory to empirics

How can we use a theoretical result like the one in equation 2 to do an empirical analysis?

A simple way to start is to derive some comparative statics and then test them in the data (you can generally do this without making any functional form assumptions, though you might need to impose some properties, such as concavity or monotonicity). For example, if you added a little bit more structure to equation 2 you could write q_{-j} as a function of the number of competitors N, and derive a comparative static that would say $\frac{\partial p}{\partial N} < 0$, i.e. the equilibrium price is decreasing in the number of firms. This is a prediction you can test in the data (provided you have some plausibly exogenous variation in N).

A more structured approach would be to take the profit function from the maximization problem, $p(q, Q) \times q - C(q)$—which you can also write in terms of price as$p \times D(p) - C(D(p))$—and estimate its empirical counterpart. To do so, you need two functions:

1. The **demand** function, $D(p)$ (or its inverse, $p(q, Q)$)
2. The **cost** function, $C(D(p))$ or $C(q)$

You should also keep in mind that to arrive at the FOC in equation 2 and any comparative static, you need to know how firms compete. In other words, you need to know how the **market equilibrium** is reached. The assumption that equation 2 is based on is that firms play a Nash equilibrium by choosing quantities simultaneously. A different set of assumptions could imply a different market equilibrium condition.

Building blocks of empirical IO models

Demand, cost, and the equilibrium conditions are the basic building blocks of an empirical IO model.

The demand function. Demand is usually micro-founded from a utility function, largely because it is rare (but not impossible! See, e.g., Einav et al., 2010) to observe the kind of random variation in prices that would allow us to simply trace demand.[1]

Most demand models estimated empirically are part of a class we call random choice utility models, where each consumer's utility is specified as the sum of a fixed and a stochastic component. For example,

$$u_{ij} = \beta_{ij} X_{ij} - \alpha p_j + \varepsilon_{ij} \tag{3}$$

Where i is the agent, j the product (or, more generally, an element of the choice set J), X_{ij} is a matrix of product/agent characteristics, p_j is price, and ε_{ij} is a random shock.

To go from this utility function to a choice, notice that

$$\mathcal{P}(i \text{ choose } j) = \mathcal{P}\left(u_{ij} \geq \max_k u_{ik}\right)$$

This probability depends on the stochastic component ε_{ij}. Most IO demand models make convenient assumptions on ε_{ij} to write down this probability analytically. The distribution that allows you to do so is the logit distribution (also known as Gumbel, or Type I generalized extreme value distribution). Under this distribution the choice probability can be written as

$$\mathcal{P}(i \text{ choose } j) = \frac{\exp(\beta_{ij} X_{ij} - \alpha p_j)}{\sum_{k \in J} \exp(\beta_{ik} X_{ik} - \alpha p_k)}$$

The cost function. The cost function, together with the demand function, determines price. How cost functions are specified depends on the problem at hand. In many cases, we can make

[1] Micro-foundation also has the advantage to allow for more flexible counterfactuals. For example, nonparametric estimation of the demand function a-la Einav et al. (2010) will let you simulate demand for existing products at prices not observed in the data, but it won't let you simulate demand for a new product—which you'd be able to do, under some conditions, if you estimated the full demand function.

simple assumptions such as constant marginal cost. In other cases, understanding the cost structure is very important. For example, Benkard (2000), focuses on learning-by-doing in aircraft production and specifies cost as a function of past production.

The equilibrium condition. The demand function and cost function come together to determine prices (and, therefore, quantity, and other market outcomes) through the equilibrium condition. The equilibrium condition specifies how the market clears. This depends on factors such as the **control variable** (e.g., what do firms set? Usually prices or quantities), and the **information structure** (i.e., what do firms know when they choose the control variable?).

To give a concrete example, in the Cournot model we discussed above, the control variable is quantity, and the information structure specifies that firms know the demand function, their own cost and that of rivals, but do not know the quantity other firms are setting (because quantity choices are made simultaneously).[2] Under these assumptions, the FOC in equation 2 is the market clearing condition that combines demand and cost to determine prices.

Empirical IO uses data to estimate these building blocks, and then leverages economic models to analyze how markets work.

To conclude this section, I want to stress that empirical IO and structural estimation are **not** the same thing. Any paper that asks an IO question and answers it using data is an empirical IO paper. Conversely, there are plenty of papers that use structural estimation to answer non-IO questions. Still, a vast amount of structural estimation techniques were developed to tackle IO questions, and a lot of modern work in empirical IO has used these techniques.

[2] In the Stackelberg model, firms move sequentially, so the firm that moves second knows the quantity chosen by the firms that moved first.

2 INTRODUCTION TO STRUCTURAL MODELING

Before jumping into how structural estimation works, it is useful to think about what structural estimation *is*. How does it differ from the more standard econometric analysis you are familiar with? What are its key principles? How *does* one build a structural model anyway? Most of what I write here is an adaptation of Reiss and Wolak (2007), which I invite you to read in its entirety.

2.1 What are structural econometric models?

Let's start with a simple example.

Suppose you have a dataset containing the winning bids, $y = \{y_1, \ldots, y_N\}$, in each of a large number of N similar auctions. You also observe the number of bidders, $x = \{x_1, \ldots, x_N\}$ in each auction. To understand the equilibrium relationship between the winning bids and the number of participants in each auction you could start by regressing y on x. Under standard statistical assumptions, this regression would deliver the best linear predictor of winning bids given the number of bidders. This is the simplest approach. If you thought the relationship could be non-linear you might use a nonparametric approach to estimate the conditional density of a winning bid given each distinct observed number of bidders x, $f(y|x)$. If variation in x arises for reasons unrelated to y, the linear or nonlinear estimate that you recover will tell you the *causal* effect of x on y. Otherwise, you may be able to use an instrument to recover the causal effect.

Other than identifying the winning bid as y and the number of bidders as x, none of the methodologies described in the previous paragraph rely on economic theory. We have not specified the nature of the auctions (e.g., sealed-bid versus open-outcry; first price versus second-price). We have not made any assumptions about bidder behavior or characteristics (e.g., risk-neutrality, expected profit maximization). You could incorporate some of these considerations as control variables, but only if data is available.

This is not an accident. By and large, reduced-form approaches require little to no economic theory. This is a feature, not a bug! It means that the result rests on very few assumptions. At the same time, lack of economic theory has two downsides. First, economic theory can clarify *why* a given relationship exists, thus providing valuable context for a result. Second, without economic theory, the estimates will usually lack external validity: in the example above, you may be able to identify the causal effect of number of bidders on the winning bid, but you wouldn't be able to predict how the winning bid will change if the number of bidders increases (or decreases) beyond the observed values of x.[3]

Two key reasons to use economic theory then, are (i) to clarify how institutional and economic conditions affect the relationship between y and x, and (ii) to provide a theoretical framework to

[3] There is a passage in Reiss and Wolak (2007) that seems to suggest that without economic theory a researcher cannot "make causal statements about an estimated relationship." *If* that's what the authors meant, I believe the statement is incorrect. You could run an experiment to identify the impact of x on y and you would identify a causal effect even with no theory at all. A more nuanced reading is that such a causal statement is slightly hollow. Without understanding the underlying mechanism of a causal effect, can we be sure that the effect would replicate under different conditions?

conduct out-of-sample counterfactuals. You can achieve (i) by pairing your analysis with a theoretical model. To achieve (ii) however, a theoretical model is not enough. You need to also have a *statistical* model—i.e., something that you can map to the data.

What would this look like in the example above? We can break down the process in two steps.

1. First, you have to specify a few features of the empirical setting. What kind of auction are you observing? Are all auctions identical or do their differ in some (observable or unobservable) ways? Are bidders aware of the number of other bidders? Are they risk-neutral?

2. Then you have to specify the statistical properties of the process that generates the data. Do bidders have common valuations? If so, what kind of distribution best describes the private valuations of the bidders?

Having made these decisions, you can then derive an expression for the probability density function of the winning bid. Reiss and Wolak (2007) gives the solution for the case of a sealed-bid auction with risk-neutral, profit-maximizing, Pareto-distributed-private-value bidders:

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

where θ_1 and θ_2 are parameters of the Pareto distribution (and x is the number of bidders). Your estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ would then be the values that maximize the likelihood function

$$\mathcal{L}\left(\left\{x_n, y_n\right\}_{n=1}^{N}, \theta\right) = \prod_{n=1}^{N} f\left(y_n|\, x_n,\theta\right)$$

What is different about this method? By finding $\hat{\theta}_1$ and $\hat{\theta}_2$, you now have the full data generating process. Under the assumptions specified in 1. and 2., you can simulate any auction and recover the expected winning bid (indeed, you can recover the entire distribution of the winning bid).[4]

This thought experiment hopefully clarifies one of the key tradeoffs between reduced-form and structural estimation. Reduced-form results have greater *internal* validity, because they rest on fewer assumptions. However, their *external* validity is limited. Conversely, structural estimation results rest on more assumptions. However, those assumptions provide a useful guide to project your results beyond what is observed directly in the data.

Summary: basics of a structural model

A structural econometric model has two components. An economic model, and a statistical model.

The **economic model** describes how the institutional setting and the agents' characteristics (i.e., the x) work together to generate the outcomes y. In the analysis above, the economic model specifies how an agent chooses their bid given their own private valuation and the number of other

[4] Notice that you could do even more and simulate results under a different type of auction too.

bidders. However, a model alone is not sufficient for an econometric analysis, for two reasons. First, because you do not observe the valuation. Second, because even if you did, economic models are deterministic. In other words, for a given x, an economic model will always return the same y. That will never happen in the data, because data is noisy, and (more importantly) because all models are wrong.

To rationalize discrepancies between the model and the data, one must add some statistical structure in the form of additional distributional assumptions and/or stochastic variables (i.e., error terms). In the example above, the statistical structure is the assumption that valuations follow a Pareto distribution (and the assumption that agents know this). Alternatively, we could have specified a model that assumed that valuations are a function of some observable bidder characteristics (e.g., the auctioned items could be oil leases, and the private valuations could be a combination of a common valuation plus observables such as distance between a bidding firm and the lease location, technologies needed to access the oil, etc.). The model would have yielded a predicted winning bid as a function of the observables x and some parameter β. To account for discrepancies between the model and the data you could assume that bidders observe the common value noisily. The combination of these assumptions is the **statistical model**.

A complete structural econometric model will return equations that make prediction about real-world objects as a function of observed quantities in the data and model parameters. In the auction example, the model predicts the probability that a given auction will result in a winning bid y as a function number of bidders and the parameters of the Pareto distribution. To estimate these parameters, we search for the values that maximize the likelihood of the observed bids. If you had used the valuation model instead, you would have generated a vector of predicted winning bids as a function of the parameters of the model and estimated β by minimizing the difference between the predicted values and the observed values.[5]

2.2 When should you consider structural techniques?

I like to tell students to use structural estimation when nothing else will work. It's a statement that sounds extreme until you realize that our ability to infer causal estimates is actually extremely limited for most economic questions. Consider the following research questions:

- Did the availability of public transportation help mitigate labor demand shocks during the Great Recession?
- How does consolidation of health insurance companies affect market outcomes?
- How do new technologies affect insurer cost?
- What makes scientists productive?

Think about how you would go about answering these questions. You can't run a randomized controlled trial for virtually any of these—how do you randomize public transportation/consolidation/

[5] I discuss the differences between these two approaches in Chapter 4.

innovation/productivity? If you are lucky, you may find a clever instrument, or a policy change that would allow you to identify some plausibly exogenous variation in your x—was there a budget surplus in some states that was spent on boosting public transportation? Can you argue that an insurance merger was driven by reasons exogenous to the outcomes you want to analyze?[6] Maybe there is exogenous variation in the exposure of insurance companies to some new technology. Even if you find an instrument or a policy change though, you will be limited by the constraints that your instrument/policy experiment imposes. Your IV might force you to limit the scope of your analysis to a subset of observations. Policy-driven variation could limit inference to a small subset of your variable's support. Plus, these approaches also require the assumption that your instrument or policy Z satisfies $\mathbb{E}[\varepsilon|\, Z] = 0$ (i.e., the exclusion restriction).

What if there is no instrument or policy change? Or what if you were interested in the potential impact of a policy that has not been implemented yet? In these situations, a structural approach can be extremely useful.

2.3 Good and bad structural estimation

Of course, not all structural models are created equal. A good structural model should be:

1. **Respectful of the economic institutions under consideration.** This means that the economic model should be tailored to reflect the specific characteristics of the setting under consideration (as much as possible). It does not mean that you should model every single detail of a market—that's impossible. But you should be thoughtful about what features to incorporate and what features to rule out ex-ante. I will discuss what this means more in detail when I present some examples of structural papers. But in general, this means trying to include all aspects of the setting that are relevant for the market outcome or policy you are studying.

2. **Flexible enough to adapt to the statistical properties of the data.** This means that your statistical assumptions should broadly reflect the patterns and summary statistics you observe in the data. This is important because structural estimation is often complicated and "black-boxy," and it is sometimes hard to understand how certain results arise from the data. Justifying your statistical assumptions by displaying summary statistics and reduced-form correlations in the data is an excellent way to support your estimation. Two common refrains you might hear at some point are: (i) that a structural paper is really a reduced-form paper plus a structural paper, and (ii) that the summary statistics section should convince readers of your main result before you even get to the structural model part.

3. **Sensitive to the nonexperimental nature of the data.** In many instances, structural estimation is chosen because it is difficult to identify a clean source of exogenous variation in the data. Ideally, even structural papers will build on a good instrumental variable or a natural

[6] For example, see Dafny et al. (2012).

Digression: identification vs. "structural identification"

At some point, you will hear the word "identification" in connection to a structural paper. It might come up as "structural identification", or as an "Identification" section in the a structural paper. People usually say that the word "identification" means different things in the context of a structural and a reduced-form empirical paper. They are right, but the difference is more in the form than in the substance.

Broadly speaking, identification in a structural paper (or **"structural identification"**) **refers to the set of assumptions under which the parameters of interest are correctly estimated**. These assumptions include (i) the specification of the economic model, including functional form assumptions, and (ii) the structure of the error term, including distributional assumptions, and assumptions about what the agents in the model know about these terms.

In a reduced-form empirical paper, identification also comes from assumptions. These assumptions are usually exclusion restrictions that rule out omitted variable bias, thus validating a natural experiment or instrumental variable approach. In a structural paper, these assumptions are expressed as properties of the econometric model, but often imply the same conditions (e.g., that agents targeted by a certain policy are equivalent to agents that were not targeted).

Another common element of the "identification" section of a structural paper is a description of the variation in the data that pins down the parameters of the model. For example, in a demand model, what pins down the price elasticity of demand might be the choices of consumers who face the same choice set, but different prices (e.g., consumers buying cars in different zip codes, or at different times of the year). Sometimes this variation is truly exogenous. Other times... not so much. In the example of consumers and cars, prices are chosen optimally by firm practicing 3rd degree price discrimination across zip-codes, so the variation is likely endogenous. Even if the variation is not entirely exogenous (and, in fact, especially in those cases), providing a clear description is very valuable.

Even when you write a reduced-form paper, you should generally perform a similar exercise, which is to discuss the hypothetical comparison that pins down the coefficient of interest in the regression. In the case of an RCT, this is straightforward: you are comparing treated to untreated units. In the case of a natural experiment the comparison is still often between treated and untreated units, but can be limited by other conditions (for example, being within a short distance of a specific threshold that triggers the treatment). Finally, in the case of an IV, this is sometimes harder to articulate, but you are still usually comparing units with high and low values of a given Z variable, and assuming that a few other things are held constant.

experiment. Still, you will likely base your model on data that was not built for the purpose of estimating your model. As a result, it is crucial to think very carefully about how the data informs your estimates. For example, can you intuitively describe what variation/correlation leads you to estimate a specific parameter? What about the sources of that variation? Are they endogenous to your outcome of interest? Why or why not?

Before you launch into a model you should think about how well you can do 1-3. If you don't think you can—perhaps because there is little economic theory on which to build, or because the vari-

ation in the data is unlikely to support clean identification of the model's parameters—then you may want to try a different approach. As you make that decision, keep in mind the following considerations:

1. Structural assumptions can substitute for data. While reduced-form nonparametric methods are desirable for their flexibility, you will need a lot of data to precisely estimate the relationship between two variables nonparametrically. Making assumptions about this relationship through an economic model can reduce the number of parameters you need to estimate and lessen the data requirements.[7]

2. Structural assumptions *can* substitute for clean identification... but probably shouldn't. Occasionally, a structural approach is used in contexts where it's not possible to identify a natural experiment or a simple IV in the data. In these situations, a researcher writing a structural model can always specify assumptions under which the model parameters are identified (see the remark on identification vs. "structural identification"). But these assumptions will not solve the underlying causal identification problem. A poorly identified structural paper will deliver poorly identified parameters. There is no getting around that. However, the exercise of specifying the assumptions required for identification is often very useful! By doing so you can figure out which ones are most unrealistic. You can then work to see if you can relax them.

3. Structural work often requires richer data. This might sound like a paradox compared to 1. What it means is that structural estimation usually requires knowing more information about the market. For a concrete example, compare two papers, both of which perform empirical exercises with the goal of estimating premiums in the ACA health exchanges: Dafny et al. (2015) and Tebaldi (2024).[8] Dafny et al. (2015) estimates a nonstructural relationship between insurance premiums and number of insurers in the market. To obtain identification it uses United's decision to not participate in any markets. For this design to work you just need to know premiums (the y variable), and number of insurers (the x variable). There are other controls and a few more bells and whistles, but that's basically it. Conversely, Tebaldi (2024) estimates a full structural model of premium-setting. On top of the data requirements for the first paper, this approach also requires information on beneficiary choices, to estimate the demand function, and cost, to estimate the cost function. Then these two objects are fed into a premium-setting game between insurers to recover a first-order condition, which finally pins down premiums as a function (among other things) of the number of competitors. In general, the reason why structural work requires richer data is that your economic model will be better if it is more realistic, and realism requires a lot of information.

[7] Note that you don't need a model to make parametric assumptions (hello, OLS), but in general it's better to have some grounding in economic theory as basis for functional assumptions.

[8] The papers have different research questions, but both estimate premiums as a function of the number of insurers competing in the market.

4. If you need to simulate something that does not already exist, you almost certainly need a structural model. This is the simplest rule. If you estimate a reduced-form model you can often run some back-of-the-envelope calculation to project your results out of sample. However, these will not be particularly realistic, and will always be limited by the structure already in place. For example, in the auction case I used as a case study, you might be able to project out the winning bid as a function of number of bidders even beyond what you observe in the data. However, you won't be able to simulate what would happen if, say, you switched from a sealed-bid auction to a open outcry auction. Also, if you project out of sample you risk being made fun of.

2.4 An example: Feng et al. (2023)

To give you a practical example of the differences between non-structural and structural estimation I will introduce one of my papers: Feng et al. (2023). The paper itself uses reduced-form methods—a pretty standard event-study analysis around a policy change. It also has one of those back-of-the-envelope calculations I made fun of two sentences ago. This paper is a good case study to think through considerations **1.**–**4.** from the previous section.

Setting All you need to know for this is that Medicaid pays for drugs using a formula. The way it works is that first, state governments reimburse pharmacies that dispense prescriptions to Medicaid beneficiaries at a price p that's reflective of list prices. Then, manufacturers send a quarterly rebate to states under a program called the Medicaid Drug Rebate Program (MDRP). The rebate size depends on a statutory formula:

$$R = N \times p \times r$$

where N is the total number of Medicaid prescriptions in a given quarter, p is the same list price used in the initial reimbursement, and r is the rebate fraction. The rebate r has two components. The first (known as the "basic rebate") equals the greater of the statutory minimum rebate and the largest discount offered to a commercial payer. The second component, the *inflation penalty,* is the difference between the current p and the inflation-adjusted p at launch. Many drugs experience price growth above the inflation rate, so the inflation penalty often makes up a significant fraction of the overall rebate.

Model The Medicaid rebate formula incentivizes firms to avoid giving discounts above the statutory minimum rebate to private payers. To see why, consider a simple model where a manufacturer bargains with a representative commercial payer over a discount rate d off an exogenous list price p_{list}. Assume for simplicity that demand from both Medicaid and commercial patients is perfectly inelastic. Denote demand in the commercial market as D_C and demand in the Medicaid market as D_M. With this notation, the Medicaid price p_M is:

$$p_M = \min\left(p^0_{\text{list}}, p_{\text{list}}\right) - p_{\text{list}} \times \max\left(r, d\right) \tag{4}$$

where p^0_{list} is the list price of the drug at launch, and r is the minimum Medicaid rebate rate.

The negotiation problem between manufacturer and payer is

$$\max_d \left[p_{\text{list}}\left(1-d\right) D_C - p_{\text{list}} \max\left\{0, d-r\right\} D_M\right]^b \times \left[\Pi\left(p_{\text{list}}\left(1-d\right)\right) - \Pi_0\right]^{1-b} \tag{5}$$

where $\Pi\left(p_{\text{list}}\left(1-d\right)\right)$ represents the net benefit of the commercial payer of acquiring the drug at price $p_{\text{list}}\left(1-d\right)$, and Π_0 is the benefit of not striking an agreement. The negative term in the payoff of the drug manufacturer reflects the impact of the MFCC. If the manufacturer agrees to a discount $d > r$, it will lose some of its Medicaid profits.

The first-order condition yields the following implicit expression for the equilibrium discount rate d^{NB}:

$$1 - d^{\text{NB}} = \begin{cases} \frac{1}{\frac{1-b}{b} \times (-p_{\text{list}}) \frac{\partial \log \Delta\Pi\left(p_{\text{list}}\left(1-d^{\text{NB}}\right)\right)}{\partial x}} & \text{if } d^{\text{NB}} \leq r \\ \frac{1}{\frac{1-b}{b} \times (-p_{\text{list}}) \frac{\partial \log \Delta\Pi\left(p_{\text{list}}\left(1-d^{\text{NB}}\right)\right)}{\partial x}} + (1-r) \times MMS & \text{if } d^{\text{NB}} > r \end{cases} \tag{6}$$

where $MMS = \frac{D_M}{D_C + D_M}$ is the Medicaid market share of the drug and $\Delta\Pi\left(x\right) \equiv \Pi(x) - \Pi_0$. Notice that the two-part FOC implies possible bunching at the r threshold.

This first-order condition implies that the equilibrium discount rate d^{NB} is weakly increasing in r, at a rate that is increasing in the Medicaid Market Share.

Decomposing the effect of the policy We now discuss how a change in the statutory minimum rebate from r to r' affects firms with varying pre-reform discount rates. Firms with initial discount $d^{\text{NB}} < r$ are unaffected because their equilibrium price does not trigger the MFCC before or after the reform. Firms with $d^{\text{NB}} \in [r < r')$ will lower their discount because they are no longer affected by the MFCC. Firms with $d^{NB} \geq r'$ will also lower their price—although by a smaller amount—because their threat point has worsened. Figure 1 displays these effects graphically.

We can derive more precise formulas for how increasing the statutory minimum rebate affects each firm by imposing a functional assumption on $\Pi\left(\cdot\right)$, i.e., the profit function of the commercial payer. We assume that

$$\Pi\left(p_{\text{list}}\left(1-d\right)\right) - \Pi_0 = \alpha\left(p_{\text{list}} - p_{\text{list}}\left(1-d\right)\right) \times D_C \tag{7}$$

This functional form implies that commercial payers earn a constant linear fraction of the "savings" they generate by negotiating rebates off of list prices. This simplification allows us to rewrite the first order condition of the model as

$$d^{\text{NB}} = \begin{cases} 1-b & \text{if } d^{\text{NB}} \leq r \\ (1-b) \times (1 - (1-r) \times MMS) & \text{if } d^{\text{NB}} > r \end{cases}$$

This new condition in turn makes it very simple to compute responses to changes in r at various

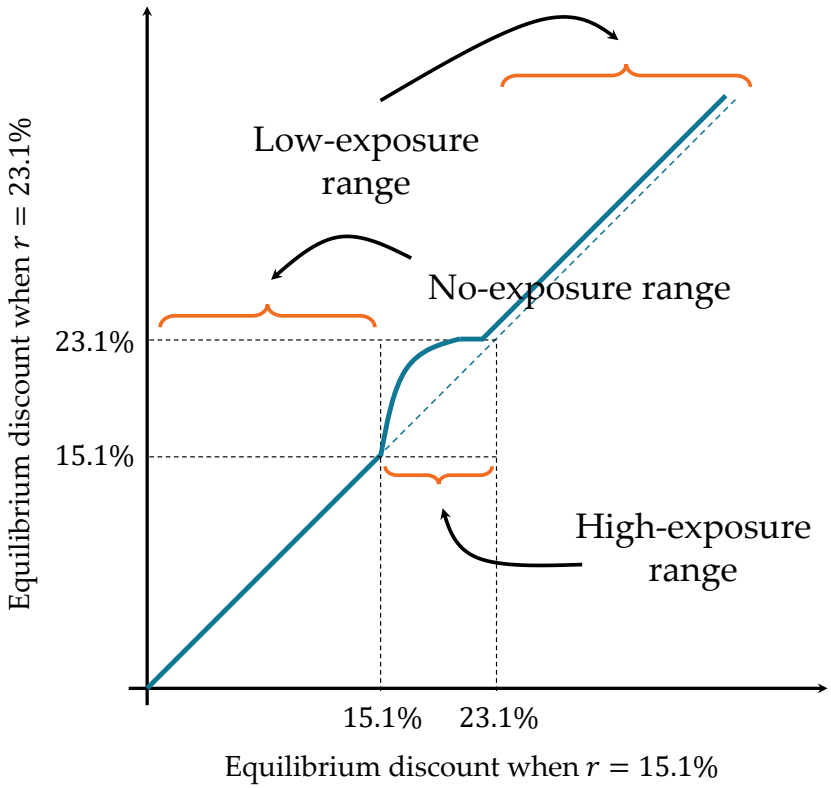

Figure 1: Effect of the Medicaid MFCC reform for drugs with different equilibrium discounts

levels of d^{NB}:

$$d' - d^{\text{NB}} = \begin{cases} 0 & \text{if } d^{\text{NB}} < r \\ (1-b)(1-r)\,MMS & \text{if } d^{\text{NB}} \in [r, r') \text{ and } d' < r' \\ r' - d^{\text{NB}} & \text{if } d^{\text{NB}} \in [r, r') \text{ and } d' = r' \\ (1-b)(r'-r)\,MMS & \text{if } d^{\text{NB}} \in [r, r') \text{ and } d' > r' \\ (1-b)(r'-r)\,MMS & \text{if } d^{\text{NB}} \geq r' \end{cases} \tag{8}$$

Using a similar logic, this model yields a prediction for the effect of removing the Medicaid MFCC altogether. This is akin to setting $r' = 1$, because in this case the MFCC can never be triggered. Let d^1 denote the equilibrium discount in this scenario. The difference between d^1 and the discount rate under floor rebate r is equal to

$$d^1 - d^{\text{NB}} = \begin{cases} 0 & \text{if } d^{\text{NB}} < r \\ (1-b)(1-r)\,MMS & \text{if } d^{\text{NB}} > r \end{cases} \tag{9}$$

Empirical analysis and results

The paper presents three sets of results. First, it estimates the effect of the policy using this regression:

$$Y_{it} = \gamma_i + \beta_1 \times \text{PostACA}_t + \beta_2 \times MMS_{i,2009} \times \text{PostACA}_t + \text{Age FE} + \varepsilon_{it} \tag{10}$$

The authors find that $\beta_2 > 0$, i.e. average discounts increases more for drugs with a high Medicaid Market Share after the reform, which confirms the predictions of the model.

Second, they compare the response of drugs whose pre-reform average discount is (i) below 15.1%, (ii) between 15.1% and 23.1%, and (iii) above 23.1%. They again find results that are consistent with the theory: the effect is largest for group (ii) and smallest for group (i).

Finally, they use the coefficients from the second regression and equation 9 to estimate d^1 (i.e., the equilibrium discount when the best-price clause is removed).

Discussion

The analysis in this paper is a useful case study to think about the differences between reduced-form methods and structural methods. To help you do this, I list some discussion topics below.

1. Why can't equation 8 be used for structural estimation?
2. What assumptions are needed to interpret the coefficient β_2 from equation 10 as causal?
3. Suppose that there was no policy reform. You could still test the predictions of the model by running a slightly modified regression:

   $$Y_{it} = \gamma_i + \beta_2 \times MMS_{i,2009} + \text{Age FE} + \varepsilon_{it}$$

   Relative to equation 10, what additional assumptions do you need to interpret β_2 causally?
4. The paper presents the results of the "no best-price" counterfactual (i.e., the estimation of d^1) as a *calibration*, without discussing assumptions about error terms. Suppose we wanted instead to treat this estimate as a serious structural counterfactual. Doing so requires explicit assumptions about the source of error terms (on top of the functional assumptions made by the authors to derive equation 9). The role of the error terms is to "justify" potential discrepancies between the predictions of the model and the data (similarly to how the residual in a linear regression justifies why any given observation might deviate from the best linear prediction). What assumptions about the error terms are the estimates of the counterfactual structurally identified? *[Hint: you need to specify (i) which variable(s) contains the error term (e.g., the discount, the Medicaid Market Share, the list price, etc.), (ii) what the source of the error is (e.g., unobserved heterogeneity, measurement error, etc.), and (iii) what firms know about the error term.]*

I also invite you to think more generally about the following issues:

- **Issue 1: structural assumptions vs. data.** *Why does the paper make additional assumptions in the section where it decomposes the effect of the policy?*
- **Issue 2: structural assumptions vs. identification.** *What is the purpose of the policy? Shouldn't you be able to identify the effect of the MFCC by simply comparing prices of drugs with different MMS?*
- **Issue 3: richness of the data.** *Why is the model in the policy effect decomposition so simple? Wouldn't it have been better to write a more realistic model?*
- **Issue 4: counterfactuals.** *Couldn't we have used the results of the first set of regressions to back out the effect of removing the MFCC?*

3 How to Build a Statistical Model: Endogeneity and Selection as Information Problems

The previous section introduced some broad conceptual ideas about structural estimation. In this section, we're going to switch gears and dive a bit more deeply into the practical side of things. You are now convinced you want to use a structural estimation approach for your next project. You have an economic model that accurately describes your setting. It's time to think about setting statistical assumptions. What do you do? This section is about that.

Statistical assumptions should help you achieve two things. First, they should allow you to come up with a **computationally feasible estimation method**. This means you want to avoid assumptions that return a non-linear minimization problem with 20,000 parameters. Second, they should **deliver an estimating equation whose coefficients are "identified"**—i.e. under the assumptions of the model, they are the true parameters of the data generating process.

The second goal has a clear equivalent in reduced form work: it's the exclusion restriction.[9] It's useful to think about how you achieve the exclusion restriction in a reduced-form setting if you are concerned that $\mathbb{E}[\varepsilon|x] \neq 0$. In this situation you have three strategies:

1. Find an instrument
2. Find a natural experiment, or some other plausibly exogenous variation
3. Add more control variables

It turns out that these three strategies are also what you use in structural estimation. The main difference is that the economic and statistical structure that the model adds can sometimes help you come up with estimation strategies that might not have been obvious otherwise.

3.1 A simple example

I will once again start with a stylized example (from Reiss and Wolak, 2007). Suppose you have cross-sectional, firm-level data on output, Q_t, labor inputs, L_t, and capital inputs, K_t. I'm going to give you a regression that relates these three variables:

$$\ln Q_{it} = \theta_0 + \theta_1 \ln L_{it} + \theta_2 \ln K_{it} + \varepsilon_{it} \tag{11}$$

Let's start with a simple question: is this a reduced-form or a structural regression? On paper, this is a reduced-form regression. However, you may notice that it looks like a log transpose of a Cobb–Douglas production function: $Q = L_t^{\theta_1} K_t^{\theta_2} \times \exp(\theta_0)$. Hence, you could interpret it as a production function. How to tell the difference? Well, the main issue is that the Cobb-Douglas production function is deterministic, whereas equation 11 has an error term, ε_{it}.

[9] The reason why the first goal generally doesn't come up in reduced-form estimation is that most reduced-form estimation function are computationally straightforward—you need millions of control variables before statistical software starts complaining about inverting the OLS matrix.

The error term is the key here. If you make no assumptions about ε_{it}, then it's possible that $\mathbb{E}[\varepsilon|x] \neq 0$, and then this is a reduced-form analysis. However, you could add some statistical assumptions to make this a structural model. The simplest thing you could do is assume that ε_{it} is an independently-distributed, mean-zero measurement error in output. Under this assumption you can recover the correct estimates of the production function of the firm, $\{\theta_0, \theta_1, \theta_2\}$. The downside is that this is probably not true.

A more realistic assumption could be to specify $\theta_{it} = \bar{\theta}_0 + \varepsilon_{it}$, i.e. that differences in firms' "productivity" levels come from an error term that we can interpret as a productivity "shock". This lets us rewrite the production function as $Q = L_t^{\theta_1} K_t^{\theta_2} \times \exp(\theta_{it})$. However, this assumption raises two issues. First, labor and capital choices will be correlated with ε_{it}—which the firm observes, even though the econometrician doesn't. This is the classic **endogeneity** concern in reduced-form applied work. In practice, this means that highly productive firms will hire more workers and invest more in capital.

The second issue is that we may not observe firms with very low draws of ε_{it}, because they have exited the market. This is a **selection** problem. We only have data for a subset of the observations, and the probability of being in the data is correlated with the realization of a stochastic variable in our model.

This example illustrates the key considerations you will face when you design the statistical part of your structural model. Some assumption (e.g., discrepancies as measurement error) make estimation way easier, but are very hard to believe. Others are more realistic, but make estimation way more difficult. Endogeneity and selection are the two chief concerns you will have to contend with as you decide which way to lean with your assumptions.

Finally, notice that both of these issues arise because **the firm knows ε_{it} before choosing whether to exit, and before choosing the inputs**, but the econometrician doesn't. This is what I mean when I say that endogeneity and selection are **information problems**. If the firm did not know ε_{it} (for example because ε_{it} is measurement error, or maybe a shock in productivity that arises only after the firm has hired worker and chosen how much "capital" to use in production), then its choice of staying in/exiting, L_t and K_t would NOT depend on ε_{it} and we would not have this problem.

There is nothing particularly special about this framing, but I personally find it is a helpful way to think about structural models. Thinking about what information an agent might have before making a decision (that an econometrician doesn't) is a very good exercise to help you figure out which potential endogeneity or selection concerns arise in your setting.

3.2 Endogeneity as an information problem (Ho and Pakes, 2014)

One of the key sources of endogeneity, which shows up in virtually every demand model, is the fact that prices are correlated with unobserved product quality.

Consider the demand model from equation 3:

$$u_{ij} = \beta_{ij} X_{ij} - \alpha p_j + \varepsilon_{ij}$$

This model is part of a class of demand models called **characteristic** models, because it treats goods as bundles of characteristics (the vector X_{ij}). These models are very popular because they let you simulate the introduction of new products (defined as a new bundle of characteristics), which is very hard to do otherwise. However, it will generally be unreasonable to assume that characteristics X_{ij} fully describe everything consumers care about. As a result, these models usually add (as one of the statistical assumptions) an unobserved product-specific quality component ξ_j. The model then becomes

$$u_{ij} = \beta_{ij} X_{ij} - \alpha p_j + \xi_j + \varepsilon_{ij}$$

This unobserved component is problematic, because p_j will generally be correlated with it. Firms know ξ_j when they choose prices, so a firm whose product has a high ξ_j will set a high price. If we don't account for this through an instrument or with some other technique, then our estimate for α will be biased.

The most conceptually straightforward way to address this endogeneity issue is with an instrument: something that affects price but is not related to product quality.[10] In this section we will instead see an example of how, purely by adding economic structure, we can shine a light on a specific endogeneity concern, and identify variation in the data that is plausibly free of such concern. This in turn lets us come up with an estimator that does not use an instrument, but still (probably) yields a well-identified parameter.

Setting

The paper we consider studies how prices and physician incentives affect the choice of hospital for privately insured patients giving birth in California. The underlying research question is whether switching from fee-for-service (FFS) to a capitated (i.e., episode-based pay) model leads to differences in physician referrals. The (informal) economic model behind this question is that physicians who are reimbursed on a FFS basis will not care about the cost a delivery, so they will send patients to the most convenient/high quality hospital. Physicians who are under a capitated payment model will earn more money if they send patients to a cheaper hospital. A big question surrounding capitated payment models is whether physicians will sacrifice quality in order to make more money.

Most papers studying provider referral choices do not include price as an explanatory variable. This probably makes sense if you are studying an FFS model, but not so much when you are studying capitation.[11] If you are *comparing* capitation and FFS, you definitely need to incorporate price into the equation.

[10] The classic instrument here would be something that affects production costs.

[11] This is also a nice example of how certain assumptions make more sense in specific settings.

In most situations, adding price into the choice/utility function is problematic because of the endogeneity issue I discussed above. In a healthcare setting, you often have to contend with two additional issues:

1. **You may not actually observe the true price.** Most data (e.g., claims) reports a "list price", but the real negotiated price between an insurer and a hospital is unobserved.
2. **The agent making the choice may not know the actual price.** This is because prices are notoriously opaque in healthcare.

We will not worry about these two issues, but they do matter, and they partially motivate the approach used in the paper. We will focus instead on the endogeneity issue.

Model

To clarify where endogeneity comes in, let's go through the model. This is a static single-agent model of hospital choice. It is common to assume that this choice is joint between patient and physician, so the choice utility function reflects both patient and physician incentives. A basic version of this utility function would look like this:

$$W_{i,\pi,h} = \theta^p_\pi p\left(c_i, h, \pi\right) + \beta X_{ih} \tag{12}$$

where

- i indexes patients, π indexes insurers, and h indexes hospitals—note that insurers matter here mostly in terms of their contract (FFS vs. capitation)
- $p\left(c_i, h, \pi\right)$ is the price that insurer π pays to hospital h for a delivery as a function of certain patient characteristics c_i. Why do characteristics matter? Because capitation payments are risk-adjusted, so, e.g. riskier deliveries receive a higher payment. With FFS, the difference arises in expectation (because a hospital will, on average, provide more services during a riskier delivery).
- θ^p_π, is an **insurer-specific price sensitivity**. This is the key parameter of interest. The reason why it's insurer-specific is that our hypothesis is that θ^p_π will be higher for patients whose insurance pays physicians using capitation (relative to insurers who pay FFS).
- X_{ih} are hospital and patient characteristics (these variables are there mostly to capture patient preferences)

This is an economic model: for given values of $p_{h,\pi}$ and X_{ih}, the model predicts exactly where each patient is going to be referred. I will also add that this is a pretty unrealistic model, because it implicitly assumes that the econometrician can observe all hospital and patient characteristics that matter for the referral choice. So the question is: how do we tweak this model to make it (i) fit the data, and (ii) more realistic?

Let's start with fitting the data because it's the simplest part. We can just add a random hospital-patient specific shock to the utility function. That's our ε_{ij}. As mention in section 1.3, if you also assume that ε_{ij} follows Type 1 GEV distribution, you can write the choice probabilities as an analytical function.

But this model is still very unrealistic. What if we don't observe everything about hospitals and patients in our X_{ij} variables? Let's start with hospital. At a minimum, it is reasonable to assume that each hospital has some quality g_h that is unknown to the econometrician, but known to patients, physicians and insurers.

With these two simple changes, we can write a slightly more reasonable model:

$$W_{i,\pi,h} = \theta_{\pi}^{p} p\left(c_i, h, \pi\right) + \beta X_{ih} + g_h + \varepsilon_{ij} \tag{13}$$

Notice that the hospital fixed effect subsumes many of the hospital characteristics that would normally go into X_{ih} (i.e., all fixed hospital characteristics), so the only things that are left in X_{ih} are hospital characteristics that change across patients. The most important such characteristic is the **distance between the patient and the hospital**.

With sufficient data, one can estimate the model in equation 13 using hospital fixed effects to control for unobservable quality differences. You may still run into issues if you have very few datapoints. For example, if you only had aggregate market shares, then estimating g_h would likely be an issue because the number of parameters would be very large relative to the number of observations leading to overfitting (this is known as the incidental parameter problem). But in this case the authors have individual-level choice data, so estimating g_h should not be an issue.

The bigger issue is whether a one-dimensional quality component is sufficient to capture everything about a hospital. One concern you might have is that hospitals are not vertically differentiated. Rather, certain hospitals are better at treating certain patients. A very complicated delivery would be best carried out in a large teaching hospital with a Neonatal Intensive Care Unit (NICU), whereas a straightforward delivery with no complications may be best carried out in a more pared down setting (because in the teaching hospital simple cases get less attention). Hence, the right way of controlling for hospital quality could be with a function $g_h\left(s_i\right)$ that specifies a different unobserved quality for a given level of a delivery "riskiness" s_i.

This specification is more realistic. However, if you want to specify $g_h\left(s_i\right)$ flexibly, you will end up with too many parameters to estimate. A potential compromise would be to specify something like

$$g_h\left(s_i\right) = g_h + \beta z_h x\left(s_i\right)$$

which allows for different quality by condition, but only as a function of observable hospital characteristics z_h (and also uses some grouping of conditions $x\left(s_i\right)$, again to reduce dimensionality). Whether this compromise is satisfying depends on whether (i) you believe that z_h captures the relevant dimensions that drive heterogeneity in quality across hospitals, and (ii) you can reliably estimate the number of parameters you have in this new model.

If we take a step back, this is a familiar problem. We don't have sufficient controls in our regression, so our coefficients are affected by omitted variable bias. One possible solution is to find an instrument or natural experiment, etc. In practice, when we are talking about hospital quality and prices, it's very hard to come up with anything of the sort.

So we are a bit stuck: to solve this endogeneity problem we need to make assumptions that are unlikely to hold.

3.3 A detour: revealed preference conditions

I am going to pause and take a quick methodological detour, which will help me introduce a solution to the problem above.

I have mentioned before that the vast majority of demand models end up adding a "logit" T1EV shock to the choice utility in order to obtain an analytical formula for market shares (if you are using aggregate data), or choice probabilities (if you are using individual-level data). In the grand scheme of things, this is a perfectly reasonable assumption that, while somewhat restrictive in terms of the shape of the error term, does simplify estimation quite a bit.

However, what could you do if you did NOT want to make this assumption?

Let's go back once again to the very first demand model I showed you in equation 3:

$$u_{ij} = \beta_{ij} X_{ij} - \alpha p_j$$

Let's say we are willing to assume that this model is correct, but we do not want to impose any distributional assumptions on the error term, other than it having mean zero. We have data on choices. What is the simplest, most basic restriction that we can impose on the model by looking at choice data?

The answer is the **revealed preference** restriction, i.e. **the observed choice is the one that maximized the agent's utility among the available options**. If agent i picked option j from a set of other options $\mathcal{J}$, then we know that

$$u_{ij} \geq u_{ik} \, \forall \, k \in \mathcal{J} \tag{14}$$

This condition looks a little bit weirder than what you may be used to because it's an inequality rather than an equation. However, this too is a restriction. Presumably, there are values of β such that

$$\beta_{ij} X_{ij} - \alpha p_j < \beta_{ik} X_{ik} - \alpha p_k \tag{15}$$

for some $k \in \mathcal{J}$. Our model can reject those values. If you understand this concept, congratulations! You know what a moment inequality is.

Now, things are not that simple. You'll noticed I dropped the logit error, but I haven't replaced it with anything else. This means that my model is deterministic, and, as we learned, that makes it unfit for estimation. If we add error terms ω_{ij} and ω_{ik} on each side of the inequality above, it

becomes more difficult to reject values of β, because we don't know the realized value of ω_{ij} and ω_{ik}.

Sometimes you can recover a version of inequality 15 that holds in expectation:

$$\mathbb{E}\left[\beta_{ij}X_{ij} - \alpha p_j + \omega_{ij}\right] > \mathbb{E}\left[\beta_{ik}X_{ik} - \alpha p_k + \omega_{ik}\right]$$

Because the error term is mean zero, it will drop out. Obviously, econometricians don't observe expectations, but can approximate them as, e.g., a sample averages:

$$\sum_{i \in \mathcal{I}_j} \beta_{ij}X_{ij} - \alpha p_j > \sum_{i \in \mathcal{I}_j} \beta_{ik}X_{ik} - \alpha p_k$$

In words, this inequality means that the sum of the utility from j of all consumers who pick j is greater than the sum of the utility from k of all consumers who picked j. As long as the number of consumers is large enough, the error term will cancel out.

Or will it? The additional assumption you need is that $\mathbb{E}\left[\omega_{ij} | \, i \text{ choose } j\right] = 0$—i.e., the expected value of the error term *conditional on choosing option* j, is equal to 0. That's a hard assumption to swallow. Agents are much more likely to pick option j when ω_{ij} is high.

Figuring out a way to cancel out ω_{ij} is hard, but notice, once again, that it's yet another version of the endogeneity problem you might have when running an OLS regression. Your x is correlated with the error term. As such, it can sometimes be solved with similar approaches (e.g., interact the expectation with an instrument Z such that $\mathbb{E}\left[Z | \, \varepsilon\right] = 0$). But what if we don't have an instrument? In those cases, sometimes, the putting some structure on the problem can help map a way forward.

3.4 Ho and Pakes (2014) II, reprise

Let's go back to Ho and Pakes (2014). We have concluded that using the standard demand estimation approach requires too many assumptions. Could revealed preferences help us?

First, let's write down a version of the choice utility function that allows for the greatest flexibility in the unobserved quality component. That version could look like this:

$$W_{i,\pi,h} = \theta_{\pi}^{p} p\left(c_i, h, \pi\right) + g_{\pi}\left(q_h\left(s\right), s_i\right) + f_{\pi}\left(d\left(l_i, l_h\right)\right)$$

where $g_{\pi}\left(q_h\left(s\right), s_i\right)$ is an insurer-specific function mapping unobserved quality to a hospital fixed effect interacted with admission "severity" s_i, and we have simplified X_{ih} to just a function $f\left(\cdot\right)$ of the distance between patient and hospital $d\left(l_i, l_h\right)$ (where l is location).[12] Notice that we have also

[12] The insurer-specific component of the hospital quality may matter if, e.g., different insurers have different contracts with hospitals, which may lead doctors to treat patients differently. Stripping all other hospital characteristics except for distance lets us simplify the notation in the derivation (which is about to get messy), but it's also quite reasonable since the new unobserved quality function $g_{\pi}\left(\cdot\right)$ likely accounts for most other variables we could include in X_{ih}.

dropped the logit shock.[13]

This model, coupled with data on where patients were referred, can yield revealed preference restrictions. We know that, for patient i, covered by insurer π, choosing to deliver in hospital h, $W_{i,\pi,h} > W_{i,\pi,k}$ for all other hospitals k that the patient could have chosen. This restriction implies

$$\underbrace{\theta_\pi^p p\left(c_i, h, \pi\right) + g_\pi\left(q_h\left(s\right), s_i\right) + f_\pi\left(d\left(l_i, l_h\right)\right)}_{W_{i,\pi,h}} > \underbrace{\theta_\pi^p p\left(c_i, k, \pi\right) + g_\pi\left(q_k\left(s\right), s_i\right) + f_\pi\left(d\left(l_i, l_k\right)\right)}_{W_{i,\pi,k}} \quad (16)$$

As written, this inequality is not particularly helpful, because we don't know anything about $g_\pi\left(q_h\left(s\right), s_i\right)$ and $g_\pi\left(q_k\left(s\right), s_i\right)$ (e.g., hospital quality specific to patient i). As a result, it's very difficult to rule out any values of θ_π^p (which is our parameter of interest). If you don't know anything about some component of the model, your best strategy might be to somehow get rid of it. But how?

Suppose there is another patient, patient i', with the same insurer as patient i, and who looks similar to patient i (formally, this means they have the same severity: $s_i = s_{i'}$), but who delivers in hospital k instead of hospital h. For patient i' the inequality condition is

$$\underbrace{\theta_\pi^p p\left(c_{i'}, k, \pi\right) + g_\pi\left(q_k\left(s\right), s_{i'}\right) + f_\pi\left(d\left(l_{i'}, l_k\right)\right)}_{W_{i',\pi,k}} > \underbrace{\theta_\pi^p p\left(c_{i'}, h, \pi\right) + g_\pi\left(q_h\left(s\right), s_{i'}\right) + f_\pi\left(d\left(l_{i'}, l_h\right)\right)}_{W_{i',\pi,h}} \quad (17)$$

What happens if we sum inequalities 16 and 17? We get

$$\begin{aligned}&\theta_\pi^p p\left(c_i, h, \pi\right) + \boldsymbol{g_\pi\left(q_h\left(s\right), s_i\right)} + f_\pi\left(d\left(l_i, l_h\right)\right) + \theta_\pi^p p\left(c_{i'}, k, \pi\right) + \boldsymbol{g_\pi\left(q_k\left(s\right), s_{i'}\right)} + f_\pi\left(d\left(l_{i'}, l_k\right)\right)\\ &> \theta_\pi^p p\left(c_i, k, \pi\right) + \boldsymbol{g_\pi\left(q_k\left(s\right), s_i\right)} + f_\pi\left(d\left(l_i, l_k\right)\right) + \theta_\pi^p p\left(c_{i'}, h, \pi\right) + \boldsymbol{g_\pi\left(q_h\left(s\right), s_{i'}\right)} + f_\pi\left(d\left(l_{i'}, l_h\right)\right)\end{aligned}$$

In this new inequality, the pairs in blue and red on opposite side of the equation are the same, and so they can be dropped, giving a simplified expression

$$\begin{aligned}&\theta_\pi^p p\left(c_i, h, \pi\right) + f_\pi\left(d\left(l_i, l_h\right)\right) + \theta_\pi^p p\left(c_{i'}, k, \pi\right) + f_\pi\left(d\left(l_{i'}, l_k\right)\right)\\ &\qquad > \theta_\pi^p p\left(c_i, k, \pi\right) + f_\pi\left(d\left(l_i, l_k\right)\right) + \theta_\pi^p p\left(c_{i'}, h, \pi\right) + f_\pi\left(d\left(l_{i'}, l_h\right)\right) \quad (18)\end{aligned}$$

Which is an inequality without unobserved heterogeneity, and where everything is either something we can see in the data, or a parameter we want to estimate. In practice, $p\left(c_i, h, \pi\right)$ (the price paid to the hospital), and $d\left(l_i, l_h\right)$ (the distance between the hospital and the patient) are observed in the data, and θ_π^p and $f_\pi\left(\cdot\right)$ are parameters to be estimated ($f_\pi\left(\cdot\right)$ is a function that you could approximate as e.g., a second-order polynomial, giving you two parameters to estimate). Intuitively, this inequality uses variation in hospital choices between identical-looking patients that live in different locations, and pay different amounts of money to hospitals.

The condition in 18 rules out values of θ_π^p that do not satisfy it. With each additional pair i and i' of patients, we can rule out more and more values of θ_π^p, until we are left with a set that's small

[13] In this context we can drop the logit shock because there is so much flexibility in the hospital quality function that we can probably fit the data (or almost fit the data) using just that function.

enough to yield an economically meaningful insight. That's what the paper does.

Discussion

How should we think about this solution? My interpretation is that it uses economic and statistical assumptions to identify variation in the data that is "less likely" to be driven by unobserved heterogeneity. Basically, the model selects certain groups of observations and compares them in a way that accounts for a host of possible confounders. In doing this selection you lose some precision:

- You don't use all the observations (there may be some patients without a counterpart)
- And you use inequality restrictions, which are less powerful than equations

But you get more robust estimates.

How does this approach differ from how you would search for a solution to an endogeneity problem in a reduced form analysis? I would argue that this approach is essentially a combination of strategies 2. and 3. (find some exogenous variation in the data and add more control variables). The paper uses the economic and statistical model to think about where the source of endogeneity is coming from (unobserved, hospital-condition specific quality). This is the "more control variables" part (they are not actually adding more control variables, but they are specifying what control variables you'd need to include). Then, they use it to figure out that there are pairs of observations in the data that, when appropriately combined, can generate plausibly exogenous variation. Fundamentally, the paper's indentification strategy assumes that, conditional on choosing the same insurer and having the same observed admission "severity," patient location and other conditions c_i (that do not enter s_i) are as good as randomly assigned. This is not foolproof. The assumptions rule out that patients may have idiosyncratic preferences for hospitals. It also requires some variation in c_i within s_i, because if $c_i = s_i$, then both patients pay the same amount, which kills the variation that identifies the paper.[14] In practice, this means that there needs to be some patient characteristic that shifts hospital payments, but does not affect hospital quality.

3.5 Selection as an information problem

I add a quick note about **selection** as a source of bias. The issue of selection arises because of endogeneity, but takes on a more severe form when the choice that a firm/agent makes affects not just the data you see, but whether you see data at all. The classic example is market entry (or exit). Examples of prediction problems that are complicated by the presence of entry/exit selection include:

- Predicting revenue for a candidate molecule undergoing clinical trials using data from similar molecules on the market

[14] Notice that if $c = s$, then $c_i = c_{i'}$ and the $\theta_\pi^p p\left(c_i, h, \pi\right)$ terms cancel out in inequality 18.

- Predicting the effect of a future merger by doing a retrospective analysis of past mergers[15]
- Evaluating the potential impact of a policy change based on data from a smaller, voluntary pilot program[16]

Some of the most important papers in IO are about dealing with the consequences of entry and exit. These include Olley and Pakes (1996) on the estimation of production functions, Ericson and Pakes (1995) and Bajari et al. (2007) on the estimation of dynamic games with imperfect competition, and Ciliberto and Tamer (2009) on estimation in the presence of multiple equilibria.

[15] Two selection effects are at play here, and they are opposite one another. The first is that firms will only attempt to merge if it is profitable for them. The second is that only mergers that do not harm consumer welfare are allowed by antitrust regulators. See, e.g., Kwoka (2014); Bhattacharya et al. (2026); Feng et al. (2026).

[16] Einav et al. (2022) shows this problem very clearly by analyzing a Medicare pilot program that was originally mandatory (and randomized), but then became voluntary after two years.

4 An Introduction to Structural Estimators

In this section I provide a brief overview of the estimation techniques that are most commonly used in structural papers. Note that this is not meant to be a comprehensive or even a rigorous introduction to structural estimation methods. I will not provide any proofs for the results I will discuss. Rather, my treatment of this material aims to provide an intuitive understanding of the way structural parameters are estimated.

4.1 Minimum distance estimators

The most common type of estimator you will encounter in structural models is the *minimum distance estimator* (MDE). MDEs are a very general class of estimators that covers virtually everything you may have seen in an econometrics course.

To start, consider a very general setup. You have an econometric model given by

$$u_n = F(x_n, \beta_0) \tag{19}$$

$$z_n = G(x_n, \beta_0) \tag{20}$$

where $\mathbb{E}[u_n \otimes z_n] = 0$.[17] We are interested in estimating certain elements of the vector β_0. To do so, we can use a sample analogue of $\mathbb{E}[u_n \otimes z_n]$ and minimize its value.

For example, suppose the data are generated by a standard linear model

$$y_n = \beta_0 x_n + \varepsilon_n$$

and let z_n be a vector of instruments—i.e., variables in the data satisfying the exclusion restriction $\mathbb{E}[\varepsilon_n z_n] = 0$ (a relevance condition—$\mathbb{E}[z_n x_n] \neq 0$—is also required for identification). Setting

$$u_n = y_n - \beta_0 x_n$$

(where $u_n = \varepsilon_n$ at the true parameter) gives the moment condition $\mathbb{E}[u_n \otimes z_n] = 0$. The resulting MDE for this setup is

$$\frac{1}{N} \sum_n z_n' \left(y_n - \hat{\beta} x_n \right) = 0 \tag{21}$$

which is the standard IV estimator. OLS is the special case where $z_n = x_n$ (the regressor used as its own instrument, valid under the exogeneity assumption $\mathbb{E}[\varepsilon_n x_n] = 0$). So OLS, 2SLS, and IV are all special cases of MDE.

The general MDE approach that you will encounter in structural papers simply consists of generating a model-predicted value for an observable data element and then minimizing the "dis-

[17] $\otimes$ here denotes the *Kronecker product*. The Kronecker product of two matrices is the block matrix obtained by multiplying all the elements of the first matrix by the second vector matrix. See https://www.statlect.com/matrix-algebra/Kronecker-product for a graphical illustration of this procedure.

tance" between the two. In the case of OLS, the data element to be matched is y_n, while the "model-predicted value" is $\beta_0 x_n$. OLS returns the values $\hat{\beta}$ that minimize the mean squared error between the two.

Most of the time, structural papers will predict data elements using relatively sophisticated models. For example, a paper may back out an expression for health plan premiums from a premium-setting game between competing insurers. Sometimes, the prediction will be a complicated nonlinear analytical expression. Other times, it may be impossible to write down an analytic expression, and the only way to generate model-predicted values is to simulate the model (that's what "simulated" estimators, like Simulated Method of Moments, do).

Another common type of estimator that falls under MDE is the Generalized Method of Moments (GMM). The main difference between GMM and MDE is that GMM has nice asymptotic properties that make calculation of standard errors easier. Hansen (1982) is the seminal paper that provides the conditions under which these properties hold. Notice however, that these properties are not necessary for getting the right estimates. In many cases you can run the same exact procedure as GMM even when the GMM conditions outlined in Hansen (1982) aren't satisfied. In these cases, bootstrapped standard errors usually work.

4.2 Maximum Likelihood Estimators

The other type of estimators you might see are Maximum Likelihood Estimators (MLEs). I am going to assume that you have seen these kinds of estimators before. If you haven't, you should first learn the basics principles behind these estimators (you can find them in pretty much any econometric textbook).

The main difference between MLEs and MDEs is that MLEs usually require additional assumptions on the error term. The only condition MDEs impose is usually that the error has mean zero (that's what equation 21 imposes). Conversely, MLEs are expressions about probabilities, and the simplest way to write down such an expression is to assume that the error term follows a specific distribution.

As an example of the differences in these two approaches, consider a simple model of firm entry. The setup we follow is similar to models presented in Bresnahan and Reiss (1990) and Bresnahan and Reiss (1991). A set of firms plays a game over two periods. In the first period, each firm decides whether to enter a market or not. In the second period, firms that entered compete in some oligopolistic game. For simplicity, we assume firms are identical, so the value of entering a market i depends only on the number of firms N_i. Under these conditions, we can write the value of entering as

$$EV\left(N_i\right) = VC\left(N_i, M_i, x_i; \theta_1\right) - FC_i \tag{22}$$

where M_i is the size of the market, x_i are controls for revenue shifters, and FC_i is the fixed cost of entering market i. The equilibrium condition is

$$VC\left(N_i, M_i, x_i; \theta_1\right) > FC_i > VC\left(N_i + 1, M_i, x_i; \theta_1\right) \tag{23}$$

How can we solve this model? For simplicity, let's assume that we have a way of recovering $VC\left(N_i, M_i, x_i; \hat{\theta}_1\right)$. Then, estimating FC_i is all that's left to do. To start, we can assume that FC_i can be written as a function of some observed and unobserved parameters, e.g.,

$$FC_i = \theta_2 z_i + \varepsilon_i \tag{24}$$

where z_i are observable market characteristics and ε_i is an unobserved, mean-zero component. If we make no other assumptions about ε_i, then we only have inequality conditions for FC_i, and this becomes a relatively challenging problem to solve. For any given market i, our condition in 23 tells us that

$$VC\left(N_i, M_i, x_i; \theta_1\right) > \theta_2 z_i + \varepsilon_i > VC\left(N_i + 1, M_i, x_i; \theta_1\right)$$

Before we use this to impose restrictions on θ_2, we need to eliminate ε_i (otherwise, if ε_i has infinite support, we can always find a value of ε_i that satisfies the constraint). Usually, the way to do so is to average across markets. In other words, since condition 23 holds for all i, we know that

$$\frac{1}{N}\sum_i VC\left(N_i, M_i, x_i; \hat{\theta}_1\right) > \frac{1}{N}\sum_i \theta_2 z_i + \varepsilon_i > \frac{1}{N}\sum_i VC\left(N_i + 1, M_i, x_i; \hat{\theta}_1\right) \tag{25}$$

Since $\mathbb{E}\left[\varepsilon_i\right] = 0$, we expect $\frac{1}{N}\sum_i \varepsilon_i \approx 0$, i.e., the error will cancel out in a large enough sample.

There are two issues with this approach. First, you may not observe data from enough markets to hope that the error term cancels out. Second, even if you do, the expression in 25 will generally only yield an identified *set*:

$$\Theta = \left\{ \theta_2 \left| \frac{1}{N}\sum_i VC\left(N_i, M_i, x_i; \hat{\theta}_1\right) > \frac{1}{N}\sum_i \theta_2 z_i > \frac{1}{N}\sum_i VC\left(N_i + 1, M_i, x_i; \hat{\theta}_1\right) \right. \right\}$$

i.e., any value of θ_2 that satisfies 25. In some cases, you may be able to aggregate different groups of markets to obtain multiple set of inequality restrictions. But there's still no guarantee that you will obtain point identification. Moreover, searching for values of $\theta_2 \in \Theta$ tends to be computationally expensive, especially if θ_2 has high dimensionality. That's because inequality restrictions cannot rely on the usual tricks that search algorithms use to find local minima of well-behaved objective functions. Most of the time, the way to find Θ is through a brute-force grid search.

What if you really need point identification or you lack computational power? A simple solution is to assume that ε follows a specific distribution $\Phi\left(\cdot; \theta_2\right)$. Then, the likelihood of observing N_i firms in market i is given by

$$\mathcal{P}\left(N_i \left| \theta_2 \right.\right) = \Phi\left(VC\left(N_i, M_i, x_i; \hat{\theta}_1\right); \theta_2\right) - \Phi\left(VC\left(N_i + 1, M_i, x_i; \hat{\theta}_1\right); \theta_2\right) \tag{26}$$

You can use this probability to create a likelihood function:

$$\mathcal{L} = \prod_i \mathcal{P}\left(N_i \left| \theta_2 \right.\right)$$

Then, $\hat{\theta}_2$ will be the value of θ_2 that maximizes $\mathcal{L}$.

The power of MLE lies in the distributional assumption. Intuitively, the assumption is reducing the possible variation in ε_i down to a small number of parameters (in many distributions, just the mean and the variance). In a sense, the assumption is a substitute for data. To see why, imagine we had data from a large number of markets and over many periods. Specifically, we can observe many instances of markets i with the same size M_k and characteristics x_k. Denote the set of such markets as $\mathcal{I}_k = \{i \left| M_i = M_k \right. \text{and } x_i = x_k\}$. Then, instead of assuming that ε_i follows a specific distribution, we can construct an empirical distribution $F_\varepsilon\left(\cdot\right)$ for $\theta_2 z_i + \varepsilon_i$ by noticing that, for all N

$$F_\varepsilon\left(VC\left(N, M_k, x_k; \hat{\theta}_1\right)\right) - F_\varepsilon\left(VC_k\left(N+1, M_k, x_k; \hat{\theta}_1\right)\right) = \frac{\sum_{i \in \mathcal{I}_k} \mathbb{I}\left(N_i = N\right)}{\left|\mathcal{I}_k\right|}$$

In other words, the probability that the value of $\theta_2 z_i + \varepsilon_i$ falls between $VC\left(N, M_k, x_k; \hat{\theta}_1\right)$ and $VC\left(N+1, M_k, x_k; \hat{\theta}_1\right)$ is given by the empirical probability that we observe N firms in markets with size M_k and characteristics x_k.

Hopefully, the tradeoff between these two approaches is clear. The first one is computationally more intensive, and requires a lot more data to yield similarly precise estimates. The second one is generally much faster, and can yield precise estimates even with a small number of datapoints. However, the second one also requires more assumptions. Over time, as computers have gotten more advanced, distributional assumptions have been increasingly viewed in a negative light (though plenty of papers still use them).

5 Bibliography

Bajari, Patrick, C. Lanier Benkard, and Jonathan Levin, "Estimating Dynamic Models of Imperfect Competition," *Econometrica*, 2007, *75* (5), 1331–1370. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2007.00796.x.

Benkard, C. Lanier, "Learning and Forgetting: The Dynamics of Aircraft Production," *American Economic Review*, September 2000, *90* (4), 1034–1054.

Bhattacharya, Vivek, Gaston Illanes, and David Stillerman, "Merger Effects and Antitrust Enforcement: Evidence from US Retail," *American Economic Review*, 2026, *forthcoming*.

Bresnahan, Timothy F., "Departures from marginal-cost pricing in the American automobile industry: Estimates for 1977-1978," *Journal of Econometrics*, November 1981, *17* (2), 201–227.

_ , "Chapter 17 Empirical studies of industries with market power," in "Handbook of Industrial Organization," Vol. 2, Elsevier, January 1989, pp. 1011–1057.

_ **and Peter C. Reiss**, "Entry in Monopoly Markets," *The Review of Economic Studies*, 1990, *57* (4), 531–553.

_ **and** _ , "Entry and Competition in Concentrated Markets," *Journal of Political Economy*, 1991, *99* (5), 977–1009.

Ciliberto, Federico and Elie Tamer, "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 2009, *77* (6), 1791–1828. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA5368.

Dafny, Leemore, Jonathan Gruber, and Christopher Ody, "More Insurers Lower Premiums: Evidence from Initial Pricing in the Health Insurance Marketplaces," *American Journal of Health Economics*, January 2015, *1* (1), 53–81.

_ **, Mark Duggan, and Subramaniam Ramanarayanan**, "Paying a Premium on Your Premium? Consolidation in the US Health Insurance Industry," *American Economic Review*, April 2012, *102* (2), 1161–1185.

Einav, Liran, Amy Finkelstein, and Mark R. Cullen, "Estimating Welfare in Insurance Markets Using Variation in Prices," *The Quarterly Journal of Economics*, 2010, *125* (3), 877–921.

_ **,** _ **, Yunan Ji, and Neale Mahoney**, "Voluntary Regulation: Evidence from Medicare Payment Reform*," *The Quarterly Journal of Economics*, February 2022, *137* (1), 565–618.

Ericson, Richard and Ariel Pakes, "Markov-Perfect Industry Dynamics: A Framework for Empirical Work," *The Review of Economic Studies*, January 1995, *62* (1), 53–82.

Feng, Josh, Thomas Hwang, and Luca Maini, "Profiting from Most-Favored Customer Procurement Rules: Evidence from Medicaid," *American Economic Journal: Economic Policy*, May 2023, *15* (2), 166–197.

_ , _ , Yunjuan Liu, and Luca Maini, "Mergers that Matter: The Impact of M&A Activity in Prescription Drug Markets," *Journal of Political Economy: Microeconomics*, 2026, *forthcoming*.

Green, Edward J. and Robert H. Porter, "Noncooperative Collusion Under Imperfect Price Information," *Econometrica*, January 1984, *52* (1), 87.

Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1982, *50* (4), 1029–1054.

Ho, Kate and Ariel Pakes, "Hospital Choices, Hospital Prices, and Financial Incentives to Physicians," *American Economic Review*, December 2014, *104* (12), 3841–3884.

Kwoka, John, *Mergers, Merger Control, and Remedies: A Retrospective Analysis of U.S. Policy*, MIT Press, December 2014. Google-Books-ID: qcYQBgAAQBAJ.

Olley, G. Steven and Ariel Pakes, "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 1996, *64* (6), 1263–1297.

Pakes, Ariel, "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, July 1986, *54* (4), 755.

Pellegrino, Bruno, "Product Differentiation and Oligopoly: A Network Approach," *American Economic Review*, 2025, *115* (4), 1170–1225.

Reiss, Peter C. and Frank A. Wolak, "Structural Econometric Modeling: Rationales and Examples from Industrial Organization," in James J. Heckman and Edward E. Leamer, eds., *Handbook of Econometrics*, Vol. 6, Elsevier, January 2007, pp. 4277–4415.

Rosse, James N., "Estimating Cost Function Parameters Without Using Cost Data: Illustrated Methodology," *Econometrica*, 1970, *38* (2), 256–275.

Rust, John, "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 1987, *55* (5), 999–1033.

Schmalensee, Richard, "Chapter 16 Inter-industry studies of structure and performance," in "Handbook of Industrial Organization," Vol. 2, Elsevier, January 1989, pp. 951–1009.

Suslow, Valerie Y., "Estimating Monopoly Behavior with Competitive Recycling: An Application to Alcoa," *The RAND Journal of Economics*, 1986, *17* (3), 389–403.

Tebaldi, Pietro, "Estimating Equilibrium in Health Insurance Exchanges: Price Competition and Subsidy Design under the ACA," *The Review of Economic Studies*, February 2024, *forthcoming*.

Tirole, Jean, *The Theory of Industrial Organization* 1988.