Lecture 23: Endogeneity I

ResEcon 703: Topics in Advanced Econometrics

Matt Woerman
University of Massachusetts Amherst

Agenda

Last time

- Rust (1987)

Today

- Endogeneity in Structural Models
- Research Design
- BLP Estimation
- Control Function

Upcoming

- Reading for next time
  - Nevo (2000) or Berry et al. (1995)
- Final exam
  - Final exam is posted, due December 19

Endogeneity in Structural Models

Endogeneity in Structural Models

So far, we have (mostly) assumed that all of our explanatory variables are exogenous

- When we talked about GMM estimation, we talked how we can use it to incorporate instruments, but I did not say much about why we would want to do so

Why is exogeneity/endogeneity so important?

- We need exogenous variation in our explanatory variables if we want to interpret our estimates as causal, rather than correlations
- But in most cases, exogenous variation is difficult to come by

Then how do we estimate causal effects?

Examples of Endogeneity

Housing choice and commute choice are correlated

- Example: people who like public transit tend to live closer to transit stations, making their transit travel time lower
- The coefficient on transit travel time will be biased upward

Price and unobserved quality are correlated

- Example: products with higher unobserved quality cost more and are preferred by consumers
- The coefficient on price will be biased downward and may even have the wrong sign

Price and unobserved marketing are correlated

- Example: large marketing campaigns may be accompanied by sales and/or increased prices
- The coefficient on price will be biased, but the direction is uncertain

Exogenous Variation and Causal Identification

How do we estimate causal effects?

Research design

- Exploit random or quasi-random variation through (natural) experiments

BLP estimation

- Use instruments to isolate exogenous variation

Control function

- Use instruments to control for endogeneity

Research Design

Credibility Revolution

The "credibility revolution" in applied microeconomics has focused on using experimental or quasi-experimental approaches to estimate causal effects

- Randomized control trials with exogenous treatment assignment
- Difference-in-differences with an exogenous policy change
- Regression discontinuity with an exogenous threshold
- Instrumental variables with exogenous instruments

Use of these methods does not guarantee exogenous variation and causal identification

- For example, if your policy change is endogenous, then the difference-in-difference estimator will be biased

But these methods can provide a framework for thinking about exogenous variation and looking for sources of (quasi-)exogenous variation to exploit

Credibility in Structural Estimation

We can combine these “treatment effect” or “reduced-form” approaches with a structural model to credibly identify structural parameters

- But the way to do so is not always obvious
- Example: How do we incorporate a regression discontinuity into a discrete choice model?

These research designs typically generate some exogenous variation, but they do not necessarily remove all of the endogeneity from our structural parameters of interest

- Example: An exogenous policy change affects the joint decision of housing and commute choices, but it does not randomly assign people to locations, so we have some exogenous variation and some endogenous variation

We can use these source of exogenous variation as instruments in the methods that follow

BLP Estimation

BLP Estimation

The context for the canonical BLP estimation approach is a random coefficients model of demand for a differentiated product using market-level data

- We want to estimate how the characteristics of products affect demand
- Price is (one of) the most important characteristics to consider
- But price is almost certainly correlated with unobserved characteristics of products

BLP (Berry et al. 1995) use instruments to isolate (supposedly) exogenous variation in price

- Including instruments in a nonlinear model using market-level data is difficult

A similar procedure can be used for endogenous variables other than price

BLP Demand Model

We have data on M markets with J_m products in market m

- One of these products can be the "outside good," or buying nothing

The utility that consumer n in market m obtains from product j is

$$U_{njm} = V(p_{jm}, x_{jm}, s_n, \beta_n) + \xi_{jm} + \varepsilon_{njm}$$

- p_{jm}: price of product j in market m
- x_{jm}: vector of non-price attributes of product j in market m
- s_n: vector of demographic characteristics of consumer n
- β_n: vector of coefficients for consumer n
- ξ_{jm}: utility of unobserved attributes of product j in market m
- ϵ_{njm}: idiosyncratic unobserved utility

Endogeneity in the BLP Demand Model

We would expect the price to depend on all attributes of a product, including those that are unobserved by the econometrician

- But if consumers also get utility from those unobserved attributes, then the price is correlated with the composite error term, $\xi_{jm} + \varepsilon_{njm}$

To solve this problem, BLP use a two-step procedure

1. Estimate the average utility for product j in market m, including observable and unobservable attributes
2. Regress this average utility value on price and other observable attributes, instrumenting for price

Utility Decomposition

Decompose the utility from observed attributes and characteristics, $V(\cdot)$, into two components (with $\bar{\beta}$ and $\tilde{\beta}_n$ similarly defined)

- $\bar{V}(p_{jm}, x_{jm}, \bar{\beta})$: component that varies over products and markets
- $\tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n)$: component that varies by consumer

Then the utility that consumer n in market m obtains from product j is

$$\begin{aligned} U_{njm} &= \bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \xi_{jm} + \varepsilon_{njm} \\ &= \left[\bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \xi_{jm}\right] + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \varepsilon_{njm} \\ &= \delta_{jm} + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \varepsilon_{njm} \end{aligned}$$

where

$$\delta_{jm} = \bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \xi_{jm}$$

This term, δ_{jm}, effectively becomes a product-market constant term that represents the average utility obtained by product j in market m

Choice Probabilities

Two distributional assumptions

- $\varepsilon_{njm} \sim$ i.i.d. type I extreme value
- $\tilde{\beta}_n$ has density $f(\tilde{\beta}_n \mid \theta)$
  - The mean is already represented by $\bar{\beta}$, so θ will often be only a variance-covariance matrix

Then the choice probabilities can be expressed as functions of δ_{jm} and $\tilde{V}$

$$P_{nim} = \int \left[\frac{e^{\delta_{im} + \tilde{V}(p_{im}, x_{im}, s_n, \tilde{\beta}_n)}}{\sum_{j=1}^{J_m} e^{\delta_{jm} + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n)}} \right] f(\tilde{\beta}_n \mid \theta) d\tilde{\beta}_n$$

We can use these choice probabilities to estimate the constant terms, δ_{jm}, and the θ parameters

- But we cannot directly estimate the $\bar{\beta}$ parameters because they are subsumed into the constant terms

Instrumenting for Price

If we assume that $\bar{V}$ is linear in parameters

$$\bar{V}(p_{jm}, x_{jm}, \bar{\beta}) = \bar{\beta}'(p_{jm}, x_{jm})$$

then we can express the constant terms as

$$\delta_{jm} = \bar{\beta}'(p_{jm}, x_{jm}) + \xi_{jm}$$

Once we have estimated the constant terms, we can regress them on prices and other attributes to estimate $\bar{\beta}$

- But price depends on ξ_{jm}, so it is endogenous
- We have to instrument for price in this regression
- Instrumenting in a linear model in easy, if we have good instruments. . .

Contraction Mapping

This estimation framework is theoretically feasible

- But we have to estimate δ_{jm} for each product and market, which can easily be 100s or 1000s (or more!) of terms to estimate

BLP developed an alternate approach that does not require estimating these δ_{jm} terms in the standard way

Their insight is that these δ_{jm} terms determine predicted market shares

- We want to find the set of constant terms that set predicted market shares equal to observed market shares

BLP show that

- For a given set of other parameters, a unique vector of δ_{jm} sets predicted market shares equal to observed market shares
- There is an iterative "contraction mapping" algorithm that recovers this unique vector of δ_{jm}

Estimation

The contraction mapping is an inner algorithm loop within the larger estimation loop

Two steps to estimate this model

1. Outer loop: search over θ to optimize the estimation objective function
    1. Inner loop: use the contraction mapping to find the constant terms, δ_{jm}, conditional on θ
    2. Use this (θ, δ) to calculate choice probabilities and then the estimation objective function
2. Estimate $\bar{\beta}$ by regressing δ_{jm} on (p_{jm}, x_{jm}) with price instruments

We can estimate this model in two different

- MSLE for step 1, 2SLS for step 2
- MSM for steps 1 and 2

Control Function

Control Function Approach

The control function approach can be thought of as the opposite of the BLP approach

- The BLP approach isolates exogenous variation
- The control function approach controls for the source of endogeneity

Why might the control function approach be better than the BLP approach?

- A control function can be used even if market shares are zero
  - The constant terms in the BLP approach are not identified for zero market shares
- A control function can control for individual-level endogeneity, rather than market-level endogeneity
  - An individual-specific constant term is not identified in the BLP approach
- The control function approach does not require the contraction mapping

Control Function Model

The utility that consumer n obtains from product j is

$$U_{nj} = V(y_{nj}, x_{nj}, \beta_n) + \varepsilon_{nj}$$

- y_{nj}: endogenous explanatory variable for consumer n and product j
- x_{jm}: vector of non-price attributes for consumer n and product j
- β_n: vector of coefficients for consumer n
- ε_{nj}: unobserved utility for consumer n and product j

The endogenous explanatory variable can be expressed as

$$y_{nj} = W(z_{nj}, \gamma) + \mu_{nj}$$

- z_{nj}: exogenous instruments for y_{nj}
- γ: parameter(s) that relates y_{nj} and z_{nj}
- μ_{nj}: unobserved factors that affect y_{nj}

Endogeneity in the Control Function Model

The utility that consumer n obtains from product j is

$$U_{nj} = V(y_{nj}, x_{nj}, \beta_n) + \varepsilon_{nj}$$

where the endogenous variable, y_{nj}, can be expressed as

$$y_{nj} = W(z_{nj}, \gamma) + \mu_{nj}$$

Two assumptions

- ε_{nj} and μ_{nj} are correlated
  - y_{nj} and ε_{nj} are also correlated, so y_{nj} is endogenous
- ε_{nj} and μ_{nj} are independent of z_{nj}
  - z_{nj} is an instrument for y_{nj}

Control Function

Decompose the unobserved utility, ε_{nj}, into two components

$$\varepsilon_{nj} = E(\varepsilon_{nj} \mid \mu_{nj}) + \tilde{\varepsilon}_{nj}$$

- $E(\varepsilon_{nj} \mid \mu_{nj})$: conditional mean of ε_{nj} (conditional on μ_{nj})
- $\tilde{\varepsilon}_{nj}$: deviations from the conditional mean

By construction, the deviations are not correlated with μ_{nj}, so they are not correlated with y_{nj}

- If we can control for the conditional mean, then we control for the source of endogeneity

We create a "control function" to control for the conditional mean

$$CF(\mu_{nj}, \lambda) = E(\varepsilon_{nj} \mid \mu_{nj})$$

- The control function is often linear, $CF(\mu_{nj}, \lambda) = \lambda \mu_{nj}$

Choice Probabilities

Substituting in the control function, the utility becomes

$$U_{nj} = V(y_{nj}, x_{nj}, \beta_n) + CF(\mu_{nj}, \lambda) + \tilde{\varepsilon}_{nj}$$

Two distributional assumptions

- $\tilde{\varepsilon}_n$ has density $g(\tilde{\varepsilon}_n \mid \mu_n)$
- β_n has density $f(\beta_n \mid \theta)$

Then the choice probabilities are

$$P_{ni} = \int \int I(V_{ni} + CF_{ni} + \tilde{\varepsilon}_{ni} > V_{nj} + CF_{nj} + \tilde{\varepsilon}_{nj} \; \forall j \neq i)$$
$$\times \, g(\tilde{\varepsilon}_n \mid \mu_{nj}) f(\beta_n \mid \theta) d\tilde{\varepsilon}_n d\beta_n$$

Estimation

Two steps to estimate this model

1. Estimate $\hat{\mu}_{nj}$ by regressing y_{nj} on z_{nj}
   - $\hat{\mu}_{nj}$ is the residual of this regression
2. Estimate (θ, λ) by MSLE using the choice probabilities

An alternative approach is to estimate all parameters simultaneously

- This approach requires that we specify the joint distribution of ε_n and μ_n, whereas the sequential method requires only the conditional distribution of ε_n given μ_n
- But if we can correctly specify this joint distribution, then the simultaneous approach is more efficient

Announcements

Reading for next time

- Nevo (2000) or Berry et al. (1995)

Office hours

- Reminder: 2:00–3:00 on Tuesdays in 218 Stockbridge

Upcoming

- Final exam is posted, due December 19