Econometrics II

Lecture 1: Identification of Causal Effects

David Schönholzer

Stockholm University

April 2, 2024

Plan for Today

1. Econometric Models
   - Parametric versus Nonparametric Models
   - Descriptive versus Causal Models
2. Identification in Econometric Models
   - Formal Definition
   - Examples
3. Potential Outcomes Framework
   - Assumptions
   - Average Treatment Effects
   - Assignment Mechanisms

Table of Contents

1. Econometric Models
   - Parametric versus Nonparametric Models
   - Descriptive versus Causal Models
2. Identification in Econometric Models
   - Formal Definition
   - Examples
3. Potential Outcomes Framework
   - Assumptions
   - Average Treatment Effects
   - Assignment Mechanisms

Econometrics: Models for Data

Every empirical economics project comes down to this:

Y_i	D_i	X_i	Z_i	...
4	0	6	1	...
1	1	3	1	...
...	...	...	...	...

Econometric models provide a framework to analyze (i.e. summarize) economic data

To this end, they posit a data generating process (DGP): $(W_i, \varepsilon_i) \sim F_\theta$

- A DGP is a joint distribution of a data row: $W_i = (Y_i, D_i, X_i, Z_i, ...)$
- The DGP also includes unobservables that are not in the data, ε_i
- We pretend all rows were drawn from this DGP with parameter/structure θ

A model is a restricted family of DGPs, i.e. we make assumptions about F_θ:

- Parametric versus non-parametric: is $\theta \in \mathbb{R}^K$ or is it a function?
- Descriptive versus causal: is θ "factual" or "counterfactual"?

Example: Parametric Model

Consider

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\
(X_i, \varepsilon_i) &\overset{\text{iid}}{\sim} N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right) \\
\theta &= (\beta_0, \beta_1, \mu_1, \mu_2, \log \sigma_1, \log \sigma_2, \sigma_{12}) \in \mathbb{R}^7
\end{aligned}
$$

Everything there is to know about the data in 7 numbers!

Example: A Semi-Parametric Model

$$
\begin{aligned}
Y_i &= X_i'\beta + \varepsilon_i \\
\mathbb{E}\left[\varepsilon_i | X_i\right] &= 0 \\
\beta &\in \mathbb{R}^K
\end{aligned}
$$

- Note we did not restrict $F_X(\cdot)$ of X_i
- Only restricted mean of $F_{\varepsilon|X}(\cdot)$ of ε_i
- Both $F_X(\cdot)$ and $F_{\varepsilon|X}(\cdot)$ potentially infinite dimensional

Another Semi-Parametric Example: Index Model

$$
\begin{aligned}
Y_i &= g\left(X_i'\beta\right) + \varepsilon_i \\
\mathbb{E}\left[\varepsilon_i | X_i\right] &= 0 \\
\beta &\in \mathbb{R}^K \\
g(\cdot) &: \mathbb{R} \to \mathbb{R} \text{ is monotone increasing}
\end{aligned}
$$

- We are interested in β
- Functions $\left\{g(\cdot), F_{\varepsilon|X}(\cdot), F_X(\cdot)\right\}$ are "nuisance" (i.e. not of direct interest)

A Non-Parametric Model

$$Y_i = g\left(X_i, \varepsilon_i\right)$$
$$X_i \perp \varepsilon_i$$
$$g\left(\cdot, \cdot\right) : \mathbb{R}^2 \to [0, 1] \text{ is monotonically increasing in both arguments}$$

- Unrestricted marginals of ε_i and X_i
- Interested in function $g\left(\cdot, \cdot\right)$ or features like

$$h(X_i) = \mathbb{E}_\varepsilon \left[g\left(X_i, \varepsilon_i\right)\right]$$

Descriptive Models

- θ is "factual": capturing moments of the data, e.g. means, correlations, etc
- Goal: imagine I got a new (D_i, X_i, Z_i). What would Y_i be? $\rightarrow$ Want $\widehat{Y}_i$
- Examples:
  1. $\theta = \mathbb{E}[Y_i|X_i]$: earnings by educational background
  2. $\theta = \text{Corr}(Y_i, X_i|G_i, R_i)$: correlation of mortality and income by gender/race
  3. $\theta = F(Y_i, X_i|Z_i)$: joint distribution of wealth of children and parents by city
- These are always valid and often interesting objects

$\rightarrow$ See much of Raj Chetty's recent work

Causal Models

- θ is "counterfactual": capturing (causal) effects of treatments on outcomes
- Goal: imagine I change $D_i = 0$ to $D_i = 1$ for some i. What would Y_i be?
  - Want to pin down θ itself rather than just construct $\widehat{Y}$
  - Interested in causal mechanisms, not descriptive facts

- When do correlations speak to causality?
  - Need an explicit counterfactual: how the world would change if I manipulate data
  - Imagine our model is $Y_i = f(X_i, U_i)$. Causal effect of changing X_i from x' to x'':

$$\Delta(x'', x') = f(x'', U_i) - f(x', U_i)$$

  - Can we change X_i without changing U_i? Beware of FUQs

Examples (we will expand on all of these):

1. $\theta = \mathbb{E}[Y_i(1) - Y_i(0)]$ the average treatment effect of a college degree
2. $\theta = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1, X_i = x]$: conditional ATE of new law on large firms
3. $\theta = F_{Y(1)}(0.5) - F_{Y(0)}(0.5)$: median treatment effect of a drug

Table of Contents

1. Econometric Models
   - Parametric versus Nonparametric Models
   - Descriptive versus Causal Models
2. Identification in Econometric Models
   - Formal Definition
   - Examples
3. Potential Outcomes Framework
   - Assumptions
   - Average Treatment Effects
   - Assignment Mechanisms

Observationally Equivalent Structures

When is a model (or its parameter/structure) identified?

A couple of preliminary definitions:

- $F_\theta(\cdot)$: distribution function implied by θ under the model

Definition (Observationally equivalent structures)

Two structures θ' and θ'' are *observationally equivalent* if

$$F_{\theta'}(\cdot) = F_{\theta''}(\cdot)$$

for any value in the domain of $F_\theta(\cdot)$

Two values of θ could produce the same data $\Rightarrow$ they are observationally equivalent

The Identified Set

What values of θ are consistent with the joint distribution of the data?

- $F_W(\cdot)$: distribution function governing observed variables

Definition (The identified set)

The *identified set* of θ is the set of observationally equivalent structures:

$$\Omega(F_W, \Theta) = \{\theta \in \Theta : F_\theta(\cdot) = F_W(\cdot)\}$$

In words, the identified set is the subset $\theta \in \Theta$ we can "isolate" with data

$\Omega(F_W, \Theta)$ could be e.g. a single point, a collection of points, an interval, or all of $\mathbb{R}^K$

Identification, Identification Strategy, and Research Design

Definition (Point identification)

A model (or equivalently its parameter) is *(point) identified* if the identified set $\Omega(F_W, \Theta)$ is a singleton, i.e. there are no observationally equivalent structures.

If we knew population (i.e. *infinite* sample size), would we be able to learn θ?

- Studies empirical implications of theoretical model
- Logically precedes question of how to estimate θ
- If θ is not identified, not worth constructing estimator!

Definition (Identification strategy or research design)

An *identification strategy* or *research design* consists of assumptions about the data and the model such that a (typically causal) parameter of interest θ is identified.

Example: Mixture Model

Consider this model with observed D_i and unobserved ε_i:

$$Y_i = (1 + D_i)\varepsilon_i, \quad (D_i, \varepsilon_i) \overset{iid}{\sim} \text{Bernoulli}(p) \times N(0, 1), p \in (0, 1)$$

Claim

The parameter p is point-identified.

Proof.

By Law of Total Probability:

$$f_Y(y) = \phi(y)(1 - p) + \frac{1}{2}\phi\left(\frac{y}{2}\right) p$$

where $\phi(\cdot)$ is pdf of Standard Normal. Solving for p we have:

$$p = \frac{f_Y(y) - \phi(y)}{\frac{1}{2}\phi\left(\frac{y}{2}\right) - \phi(y)}$$

Example: Additive Model

Consider the model:

$$Y_i = U_i + V_i, \quad (U_i, V_i) \overset{iid}{\sim} N\left(\begin{pmatrix}\alpha \\ \beta\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right)$$

Claim

(α, β) is not point-identified. However, $\alpha + \beta$ is.

Proof.

Taking expectations, $\mathbb{E}[Y_i] = \alpha + \beta$, so the sum is identified.
To show that (α, β) are not separately identified:

- Define $\tilde{\alpha} = \alpha + x$ and $\tilde{\beta} = \beta - x$ for some $x \in \mathbb{R}$
- Then $F_{(\alpha,\beta)}(y) = F_{\left(\tilde{\alpha},\tilde{\beta}\right)}(y)$ for all $y \in \mathbb{R}$

□

Example: Nonlinear Model

Consider the model

$$Y_i = (X_i - \theta)^2 + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0$$

Claim

The identified set is $\Omega(F_Y, \Theta) = \left\{ \mathbb{E}[X_i] \pm \sqrt{\mathbb{E}[Y_i] - Var(X_i)} \right\}$.

Proof.

Taking expectations and solving for zero, we get:

$$\mathbb{E}[Y_i] - \mathbb{E}\left[X_i^2\right] + 2\theta\mathbb{E}[X_i] - \theta^2 = 0$$

Which yields $\Omega(F_Y, \Theta)$ as the solutions for θ (using quadratic formula). □

Table of Contents

1. Econometric Models
   - Parametric versus Nonparametric Models
   - Descriptive versus Causal Models
2. Identification in Econometric Models
   - Formal Definition
   - Examples
3. Potential Outcomes Framework
   - Assumptions
   - Average Treatment Effects
   - Assignment Mechanisms

Potential Outcomes Framekwork

- Statistics approach to causality (Neyman-Rubin) ("atheoretical" counterfactuals)
- Potential outcome $Y_i(D_i)$: outcome if D_i exposed to $D_i \in \{0, 1\}$
- Typically binary D_i but could be richer
- Example: Drug trial
  - D_i: 1 if treated, 0 if placebo
  - $Y_i(1)$: health if treated
  - $Y_i(0)$: health if placebo
- "Fundamental problem of causal inference" (Holland 1986):

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

$\rightarrow$ can only observe outcome Y_i under *one* treatment status
$\rightarrow$ can never observe individual causal effects $\tau_i = Y_i(1) - Y_i(0)$

Design-Based versus Model-Based Identification

Identification can be either design-based or model-based

1. Design-based: D_i is random(ized), conditional on $(Y_i(0), Y_i(1))$
   - Randomized control trials
   - Instrumental variables
   - Some event study designs
2. Model-based: $(Y_i(0), Y_i(1))$ is random, conditional on fixed D_i
   - Basic difference-in-differences design
   - Sharp regression discontinuity design

Useful distinction to understand the identification logic of an approach

SUTVA

- When are potential outcomes well defined?
- Typically need Stable Unit Treatment Value Assumption (SUTVA, Rubin 1986)

Definition (SUTVA)

A treatment satisfies SUTVA if

1. *No hidden treatment variation*: treatment has consistent effect on unit i
   $\rightarrow$ e.g. if $D_i \in 0, 1$, it cannot be that $D_i = 2$
2. *No interference*: my outcome only depends on my own treatment status:

$$Y_i(d_1, ..., d_N) = Y_i(d_i)$$

A Finite Population Example

i	D_i	$Y_i(0)$	$Y_i(1)$	Y_i
1	1	3	2	2
2	0	4	5	4
3	0	1	4	1
4	1	3	6	6
5	0	5	5	5
6	1	4	3	3

- Causal inference involves "imputing" missing potential outcomes
- Classic assumption: potential outcomes are missing at random (MAR):

$$D_i \perp \{Y_i(0), Y_i(1)\} \quad \forall i$$

- Conditional independence assumption (CIA): $D_i | X_i \perp \{Y_i(0), Y_i(1)\}$

PO Identification Example: Average Treatment Effect

Consider the model

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0), \quad D_i \in \{0, 1\}$$
$$\mathbb{E}[Y_i(d)|D_i] = \mathbb{E}[Y_i(d)] \text{ for } d \in \{0, 1\}$$

Claim

The average treatment effect $\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$ is identified.

Proof.

Note that

$$\begin{aligned}
\tau_{\text{ATE}} &= \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right] \\
&= \mathbb{E}\left[Y_i(1)|D_i = 1\right] - \mathbb{E}\left[Y_i(0)|D_i = 0\right] \quad \text{(due to mean indep. assumption)} \\
&= \mathbb{E}\left[Y_i|D_i = 1\right] - \mathbb{E}\left[Y_i|D_i = 0\right]
\end{aligned}$$

Selection Bias

- A similar argument shows *average treatment effect on the treated* (ATT)

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$$

  is identified if the model includes $\mathbb{E}[Y_i(d)|D_i] = \mathbb{E}[Y_i(d)]$

- Now suppose we lack this mean independence assumption
- Then identification fails because of *selection bias*:

Definition (Selection Bias)

The gap between ATT and difference between treatment and control means is:

$$\mathbb{E}\left[Y_i|D_i = 1\right] - E\left[Y_i|D_i = 0\right] = \tau + \underbrace{\mathbb{E}\left[Y_i(0)|D_i = 1\right] - \mathbb{E}\left[Y_i(0)|D_i = 0\right]}_{\text{Selection bias}}$$

Assignment Mechanism

- Whether MAR or CIA holds depends on the *design* $\{D_i\}_{i=1}^N$
- The design of a setting can be described through the *assignment mechanism*
- What determines why some units are treated and others not?
- Let $\mathbf{D} = (D_1, ..., D_N)'$ and $\mathbf{Y}(\mathbf{d}) = (Y_1(d), ..., Y_N(d))'$ for $d \in \{0, 1\}$

Definition (Assignment Mechanism)

For a population of N units, it is a function $P(\mathbf{D}|\mathbf{X}, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}))$ that satisfies

$$\sum_{\mathbf{D} \in \{0,1\}^N} P(\mathbf{D}|\mathbf{X}, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})) = 1$$

for all $\mathbf{X}$, $\mathbf{Y}(\mathbf{0})$, and $\mathbf{Y}(\mathbf{1})$ where $\mathbf{D} \in \{0, 1\}^N$ is all possible treatment assignments.

Unit Assignment Probability and Propensity Score

Definition (Unit assignment probability)

Unit i has the following probability of being treated for all $\mathbf{X}$, $\mathbf{Y}(\mathbf{0})$, and $\mathbf{Y}(\mathbf{1})$:

$$p_i(\mathbf{X}, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})) = \sum_{\mathbf{D}: D_i = 1} P(\mathbf{D}|\mathbf{X}, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}))$$

Definition (Propensity score)

The propensity score at $X_i = x$ is for all $\mathbf{X}$, $\mathbf{Y}(\mathbf{0})$, and $\mathbf{Y}(\mathbf{1})$:

$$e(x) = \frac{1}{N(x)} \sum_{i: X_i = x} p_i(\mathbf{X}, \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}))$$

where $N(x) = \#\{i = 1, ..., N | X_i = x\}$

Example: Completely Randomized Experiment

- Let us return to the finite-sample example
- Completely randomized experiment with $n_1 < N$ treated units:

$$P\left(\mathbf{D}|\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})\right) = \binom{N}{n_1}^{-1}$$

  if $\sum_{i=1}^{N} D_i = n_1$ and $P\left(\mathbf{D}|\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1})\right) = 0$

- Treatment assignment with $N = 3$ and $n_1 = 2$:

Assignment number:	1	2	3	4	5	6	7	8	p_i	$e(x)$
$i = 1$	0	1	0	0	1	1	0	1	$\frac{2}{3}$	$\frac{2}{3}$
$i = 2$	0	0	1	0	1	0	1	1	$\frac{2}{3}$	$\frac{2}{3}$
$i = 3$	0	0	0	1	0	1	1	1	$\frac{2}{3}$	$\frac{2}{3}$
$P(\mathbf{D}\|\mathbf{X}, \mathbf{Y}(\mathbf{1}), \mathbf{Y}(\mathbf{0}))$	0	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0		