SELECTION ON OBSERVABLES

Ivan A. Canay
Northwestern University[1]
ECON 481-3

1. Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. The Review of Economics and Statistics 86: 4-29
2. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

[1]Today's class is based on Alex Torgovitsky's notes. I'd like to thank him for kindly sharing them.

Topics of Part I

- Lec I: Selection on Observables
  1. Potential Outcomes vs Latent Variables
  2. Causal Inference
  3. Selection Bias
  4. Selection on Observables & Selection on Prop. Score
- Lec II: Roy Models and LATE
  1. The role of heterogeneity
  2. Multiple instruments, Covariates, and Abadie's κ
- Lec III: Marginal Treatment Effect
  1. Parameters as functions of MTEs
  2. Policy Relevant Treatment Effects
- Lec IV: Extrapolations
  1. Semi-Parametrics MTEs
  2. Weights for Target Parameters

Motivation

Counterfactual Questions

- What would happen if a job training program were expanded? [Labor]
- What would happen to prices/welfare if two firms merged? [IO]
- What would different monetary policy do to real output? [Macro]
- What effect would this medication have on heart disease? [Biostatistics]
- What will happen to global temps if emissions decrease? [Climatology]

Causal Inference

- Thinking about a counterfactual requires thinking about causality
- Theory alone (might) tell us the direction of causality
  $\Rightarrow$ Even when it does, it will rarely tell us the magnitude
- Causal inference uses data to address counterfactuals

Potential Outcome Notation

- Also known as the Neyman-Fisher-Roy-Quandt-Rubin causal model

Notation

- $\mathcal{D}$ is a mutually exclusive and exhaustive set of states ("treatments")
  e.g. training/no training $\mathcal{D} = \{0, 1\}$, prices $\mathcal{D} = [0, \infty)$, etc.
- For each $d \in \mathcal{D}$ there is a potential outcome Y_d (a random variable)
- Y_d is what *would* have happened if the state were d

- **Observed**: the actual state, a random variable $D \in \mathcal{D}$
- **Observed**: outcome Y, related to potential outcomes as

$$Y = \sum_{d \in \mathcal{D}} Y_d I\{D = d\} = Y_D \,.$$

$Y = Y_D$ is observed, but Y_d for $d \neq D$ are **unobserved**

What Do We Want to Measure?

- We are interested in **counterfactuals**, Y_d for $d \neq D$
- These variables capture the "what if" aspect of causality
- Since they are random variables, they can be summarized in many ways
- That is, there are many possible parameters of interest

Example (Program Evaluation)

- Suppose $d \in \{0, 1\}$ indicates participation in a job training program
- Y is a scalar labor market outcome such as earnings
- If $D = 1$ we observe Y_1 (but not Y_0) and if $D = 0$ we observe Y_0
- There are many possible questions one could ask:
  - What would be average earnings if everyone were trained, i.e. $E[Y_1]$?
  - What is the average effect of the program, i.e. $E[Y_1 - Y_0]$?
  - What about only for those who are trained, i.e. $E[Y_1 - Y_0 | D = 1]$?

Latent Variable Notation

- Many empirical models in economics look like a special case of:

$$Y = g(D, U) \; ,$$

  where g is a function and U are unobservable variables

- A **causal** interpretation of this model is implicitly saying:

$$Y_d = g(d, U) \text{ for every } d \in D$$

- This could impose assumptions, depending on what g and U are

Warning

- Some are dogmatic about potential outcome vs. latent variable notation
- Often follows some field-specific social norms, e.g. labor vs. IO
- **Remember**: It's just notation - use the above to translate

Outline

1. Potential Outcomes and Latent Variables
2. **Random Assignment: The fundamental Problem of Causal Inference**
3. Selection Bias and Selection on Observables
4. The Role on the Propensity Score
5. Final Remarks on Selection on Observables

Random Assignment

- **Random assignment** is the assumption that $\{Y_d : d \in \mathcal{D}\} \perp\!\!\!\perp D$
- Under random assignment, the distribution of Y_d is point identified,

$$\underbrace{F_d(y)}_{\textit{parameter}} = P\{Y_d \leq y\} \overset{(*)}{=} P\{Y_d \leq y | D = d\} = \underbrace{P\{Y \leq y | D = d\}}_{\textit{observed}},$$

  where $(*)$ follows from random assignment.
- Any parameter that is a function of $F_d : d \in \mathcal{D}$ is also point identified
- Intuition: conditioning on treatment does not change potential outcomes

$\Rightarrow$ No self-selection, sorting, correlated observables/unobservables, etc.

Parameters

Common parameters of interest with binary D

- Average treatment effect (ATE): $E[Y_1 - Y_0]$
- Average treatment on the treated (ATT): $E[Y_1 - Y_0|D = 1]$
- Quantile treatment effect (QTE): $Q_{Y_1}(t) - Q_{Y_0}(t)$ for some $t \in (0, 1)$
- QTE on the treated/untreated (QTT/QTU) defined analogously

- All point identified under random assignment
- Moreover, $ATE = ATT = ATU$, and $QTE = QTT = QTU$
- Nothing systematically different about treatment/control groups
- If D is multivalued or continuous, conventions are less established

Causal Inference: Problem

The Problem

- Even with random assignment, joint dist's, $P\{Y_1 \leq y_1, Y_0 \leq y_0\}$, **are not** point id'd:
- Sometimes called the fundamental problem of causal inference
- Intuition: we never see both Y_0 and Y_1 for anyone

Implications

- Most features of $Y_1 - Y_0$ are not point identified
- Even with random assignment $\Rightarrow$ so without it as well
- We might care about the proportion of individuals who are hurt:

$$P\{Y_1 \leq Y_0\} \rightarrow \text{ not point identified!}$$

- Nor are the quantiles of $Y_1 - Y_0$
- **Note**: Quantile treatment effect (QTE) vs. quantile of the treatment effect

Random Assignment and Covariates

Role of Covariates

- Suppose we regress Y on D and predetermined X
- D is randomly assigned, so should be uncorrelated with X
  $\rightarrow$ Common practice to check this as a "balance test"
- Also means variation in coefficient on D will **go down**
- How much depends on how much X and Y are correlated

Example

- Y is cholesterol after the experiment
- D is a drug intended to reduce cholesterol
- X is your cholesterol in the past, before the experiment
- X probably explains a lot of the variation in Y
- Controlling for X reduces residual variation in Y, but not D
- This allows one to estimate the effect of D more precisely

Scope of Random Assignment

When is random assignment a good assumption?

- Typically, settings where agents have no control over D
- Less likely: Agents choose D without considering $\{Y_d : d \in \mathcal{D}\}$
- Randomized controlled experiments are the leading case
- Random assignment is rarely compelling with observational data
- When agents can control D, we typically expect **selection**

- Random assignment leads to high internal validity
- The phrase "gold standard" is often used in biostatistics
- In practice, researchers rarely "flip a coin" (c.f. CAR)
- Random assignment often comes along with lower external validity

Outline

1. Potential Outcomes and Latent Variables
2. Random Assignment: The fundamental Problem of Causal Inference
3. **Selection Bias and Selection on Observables**
4. The Role on the Propensity Score
5. Final Remarks on Selection on Observables

Selection

Formal definition

- There is selection into the treatment state D if

$$Y_d | D = d \text{ is distributed } \textbf{differently} \text{ from } Y_d | D = d' \text{ for } d \neq d'$$

- Expected to occur if agents choose D with knowledge of $\{Y_d : d \in \mathcal{D}\}$

Selection is common

- Particularly concerning if you are trained in neoclassical economics
- Optimization: Agents choose a job training program $(D \in \{0, 1\})$ to maximize utility
- Utility: incorporates expected future earnings (Y_0, Y_1)
- Agents who choose job training might do so because of low Y_0
- Alternatively, might choose $D = 0$ because of high Y_0

Selection Bias

- Consider the simple treatment/control mean contrast under selection
- This contrast would be the ATE under random assignment
- Decompose the contrast into a causal effect and selection bias:

$$E[Y|D=1] - E[Y|D=0] = \underbrace{(E[Y_1|D=1] - E[Y_0|D=1])}_{ATT} + \underbrace{(E[Y_0|D=1] - E[Y_0|D=0])}_{\textbf{selection bias}}$$

- **First term:** causal effect for those who were treated
- **Second term:** how the treated would have been different anyway
- Effects could cancel out: ATT is $(+)$ while selection bias is $(-)$
  $\Rightarrow$ Job training program, drug for a lethal disease, etc

Selection on Observables

- A simple relaxation of random assignment is selection on observables
- Suppose that we observe (Y, D, X) where X are covariates
- The **selection on observables** assumption is that

$$\{Y_d : d \in \mathcal{D}\} \perp\!\!\!\perp D \mid X\,.$$

- Says: Conditional on X, treatment is as-good-as randomly assigned
- Other terms: unconfoundedness, ignorable treatment assignment
- Underlies causal interpretations of linear regression

Identification Argument

- Conditional version of random assignment:

$$\begin{aligned} F_d(y|x) &= P\{Y_d \leq y|X = x\} \\ &= P\{Y_d \leq y|D = d, X = x\} \\ &= P\{Y \leq y|D = d, X = x\} \end{aligned}$$

- Second equality requires **overlap**: $P\{D = d|X = x\} > 0$
- Integrating over x, one can point identify the marginals

$$F_d(y) = P\{Y_d \leq y\} = E[P\{Y_d \leq y|X = x\}] = E[P\{Y \leq y|D = d, X = x\}]$$

Identification of Mean Contrasts

- Suppose $D \in \{0, 1\}$ is binary - by far the most common case
- Using essentially the same argument as on the previous page:

$$ATE = E[E[Y_1|X] - E[Y_0|X]] = E[E[Y|D = 1, X] - E[Y|D = 0, X]]$$

- Similar expression for the ATT has an important difference:

$$ATT = E[Y|D = 1] - E[E[Y|D = 0, X|D = 1]]$$

- Helps in estimation since **only one** conditional expectation (more later)
- Note that only mean independence is needed for these arguments

$$E[Y_d|D = 0, X] = E[Y_d|D = 1, X]$$

- Difficult to think of arguments for means (without full) independence

Controlling on too much

Example

- Suppose that $(Y_0, Y_1) \perp\!\!\!\perp D | X$
- Let X_2 be a subset of X
- Let X_1 be a subset of X_2
- So controlling on X is the most information, and X_1 is the least
- Suppose however that we only have X_2 (hence X_1) in our data
- Perhaps surprisingly, using X_2 can introduce more bias than using X_1
- That is: adding information $(X_2 \textit{ vs } X_1)$ **need not** reduce bias
- If $X_2 = X$, then it does, but not more generally
- Point is not well-appreciated in applied work. But should be concerning

Outline

1. Potential Outcomes and Latent Variables
2. Random Assignment: The fundamental Problem of Causal Inference
3. Selection Bias and Selection on Observables
4. **The Role on the Propensity Score**
5. Final Remarks on Selection on Observables

The Propensity Score

Definition

- Binary treatment case: $D \in \{0, 1\}$
- $p(x) = P\{D = 1|X = x\}$ is called the **propensity score**
- Let $P = p(X)$ be the random variable $P\{D = 1|X\}$

Rosenbaum and Rubin (1983) sufficiency argument

- Selection on observables implies $(Y_0, Y_1) \perp\!\!\!\perp D \mid P$:

$$
\begin{aligned}
P\{D = 1|Y_0, Y_1, P\} &= E\Big[P\{D = 1|Y_0, Y_1, P, X\}|Y_0, Y_1, P\Big] \\
&= E\Big[P\{D = 1|Y_0, Y_1, X\}|Y_0, Y_1, P\Big] \\
&= E\Big[P\{D = 1|X\}|Y_0, Y_1, P\Big] \\
&= E\Big[p(X)|Y_0, Y_1, P\Big] = P
\end{aligned}
$$

- Implication: we can condition on P instead of X

Propensity Score Weighting

- Using p, we can write the ATE as a weighted average of Y

$$ATE(x) = E\left[\frac{Y(D - p(x))}{p(x)(1 - p(x))}\middle|X = x\right]$$

- Average over X to point identify

$$ATE = E\left[\frac{Y(D - p(X))}{p(X)(1 - p(X))}\right]$$

- Similar expressions can be derived for ATT,

$$ATT = E\left[\frac{Y(D - p(X))}{P\{D = 1\}(1 - p(X))}\right]$$

- Derivations are straightforward (see Pset)

Different Identification arguments

Summary

- Three different constructive identification results for the ATE
  $\Rightarrow$ Match on X, match on P, weight using p
- Each one shows that ATE is point identified
- And they are all derived under the **same assumptions**

So why have three?

- Identification arguments directly inform the construction of estimators
- Different arguments suggest different estimators
- In general, these different estimators may have different properties
- In selection on observables this is definitely true
  $\Rightarrow$ Subtle differences in efficiency, rates of convergence
- More importantly, differences in finite sample performance

Multivalued Treatments

- Many interesting counterfactual states are **multivalued**
- The main identification arguments mostly remain the same/similar
- Some details (Imbens, 2000) regarding the (generalized) propensity score
- However, the literature is overwhelmingly about $D \in \{0, 1\}$
- Imbens and Rubin (2015 book, 650 pages) exclusively discuss this case!
- The reason seems (to me) to be mostly sociological
- Nonparametric methods are highly valued by those in this literature
- With $D \in 0, 1$ there is only non-parametric (in D)
- If $D \in \{0, 1, 2\}$, then one needs to make a choice:
  (a) Make a (potentially wrong) functional form assumption
  (b) Remain non-parametric - basically reduces back to the binary case
- Community is against the first option, and second has low payoff for theory work.

Outline

1. Potential Outcomes and Latent Variables
2. Random Assignment: The fundamental Problem of Causal Inference
3. Selection Bias and Selection on Observables
4. The Role on the Propensity Score
5. **Final Remarks on Selection on Observables**

Criticisms of Selection on Observables

Inherent unobservables

- Selection on observables can be difficult to believe in economics

$\rightarrow$ Inherent unobservables: preferences, private info, expectations, . . .

- Observationally identical people behave differently due to . . . a coin flip?

Controlling for more: not a solution

- Often argued that large X makes selection on observables more likely
  $\Rightarrow$ This is, of course, not necessarily true - we saw this earlier
- Even if it were, still raises an uncomfortable friction with overlap
  $\Rightarrow$ If we could perfectly explain D with X then $P\{D = 1|X\}$ would be 0 or 1

Better methods for choosing observables will not solve this

- Selection on observables is seeing a resurgence with machine learning
- Fancier methods, but the identifying assumption is still the same
- Bias/variance trade-off is not the first-order issue here

Angrist (1998)

Misfit selection on observables

- Well-known economic application of selection on observables:
  - Y is a labor market outcome (employment/earnings)
  - $D \in \{0,1\}$ is veteran status (participation in the military)
  - X are socioeconomic variables (race, year, schooling, AFQT, age)
- **Assumption:** given X, military participation as-if randomly assigned
- Observationally similar people randomly join the military?!?
- Ignores first-order issues such as outside employment options
  $\Rightarrow$ Also unobservable screening factors (fitness, interpersonal skills)
- These are unobserved and **inherently unobservable**

Allowing for selection on unobservables

- Most applied microeconomists seem to share this skepticism
- This motivates the other methods that we will discuss in the course
- All use different arguments to allow for **selection on unobservables**