Lecture 6: Introduction to Causal Inference

Hammad Shaikh

Motivation

- Economic questions can often be simplified to studying the causal relationship between two variables
- Correlation between the variables is a good starting point
  - Examining scatter plot can be misleading
- Anecdotal evidence is common and arguably most influential
  - Not scientific!
- Econometrics: study causal relationship between economic variables using statistical methods

Framework for Causal Inference

- Y_i = Outcome of interest (wages)

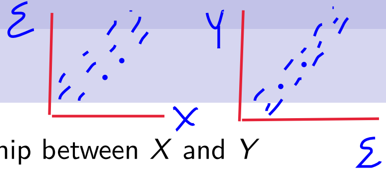
- T_i = Treatment of interest (attend college)

$T_i \in \{0,1\} \Rightarrow T_i = I(i \text{ attend college}) = \begin{cases} 1, \text{true} \\ 0, \text{false} \end{cases}$

- Individual causal effect: $Y_{i,T_i=1} - Y_{i,T_i=0}$
  - Not possible to compute

- Average treatment effect: $E(Y_{i,T_i=1} - Y_{i,T_i=0})$
  - Estimate ATE using $\underbrace{\bar{Y}_{T=1} - \bar{Y}_{T=0}}_{\text{Biased from selection into } T\in\{0,1\}}$

Problem With Using Scatter Plots



- Scatter plots represent the relationship between X and Y
  - Example: X = Years of Education, Y = Hourly Wage

ε = ability, motivation

- Simple linear regression: $Y_i = \beta X_i + \epsilon_i$

$\text{Corr}(\varepsilon, X) \neq 0$ problem

- Argue OLS estimator is biased if ϵ related to X

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \hat{\beta}_1 = \frac{\Delta y}{\Delta x} = \frac{\Delta y_x + \Delta y_\varepsilon}{\Delta X}$$

$$= \underbrace{\frac{\Delta y_x}{\Delta X}}_{\beta} + \underbrace{\frac{\Delta y_\varepsilon}{\Delta X}}_{\text{bias}}, \quad \varepsilon = \text{ability}, \ \Delta y_\varepsilon > 0$$

$$\Delta X \begin{cases} \Delta y_x \\ \Delta\varepsilon \rightarrow \Delta y_\varepsilon \end{cases}$$

Randomized Control Trials

Sample

Control

Treatment

- Experiments are ideal to study causal relationships

- Units are randomly assigned the treatment variable
  - Binary treatment: T = 0 (control) and T = 1 (treatment)

$\hookrightarrow \varepsilon \perp T \Rightarrow E[\varepsilon | T] = E[\varepsilon] = 0$

- Control group (T = 0) and treatment group (T = 1) are statistically identical prior to treatment assignment

- Unbiased estimate of ATE is $\bar{Y}_{T=1} - \bar{Y}_{T=0}$
  - No selection bias as now $\epsilon \perp\!\!\!\perp T$

$$E[\varepsilon | T=1] - E[\varepsilon | T=0] = 0$$

Problem With Experiments

- Experiments are expensive or unethical
    - Expensive to hire teachers
    - Unethical to decide whether someone goes to college

- Difficult to conduct large scale experiments
    - Hard to generalize results from small sample studies

- Need to obtain causal connection from observational data
    - Econometrics can help!

No control

Multiple Regression Review

Treat Control

Same $X_{2i}, \ldots, X_k$

- $Y_i = \beta_0 + \beta_1 T_i + \underbrace{\beta_2 X_{2i} + \ldots + \beta_k X_{ki}}_{\text{Control for selection into treat / control}} + \epsilon_i$

- Control for variables related to both Y_i and T_i

  Pre-treatment variables

- Selection on observables: T_i is essentially random after controlling for $X_{2i}, \ldots, X_{ki}$
  - Conditional independence assumption (CIA): $T_i \perp\!\!\!\perp \epsilon_i$ after accounting for $X_{2i}, \ldots, X_{ki}$

- Problem: CIA usually doesn't hold in practice

  ↳ Some variables are hard to measure

  ε = ability & motivation

Causal Inference from Observational Data

- Question: Is academic probation effective in encouraging students to continue with university and improving their performance?

Y = Dropout from univ.

- Institutional background: Academic probation is CGPA < 1.5

$$ProbStatus_i = I(iGPA < 1.5)$$

- How to determine causal effect of probation on outcomes?

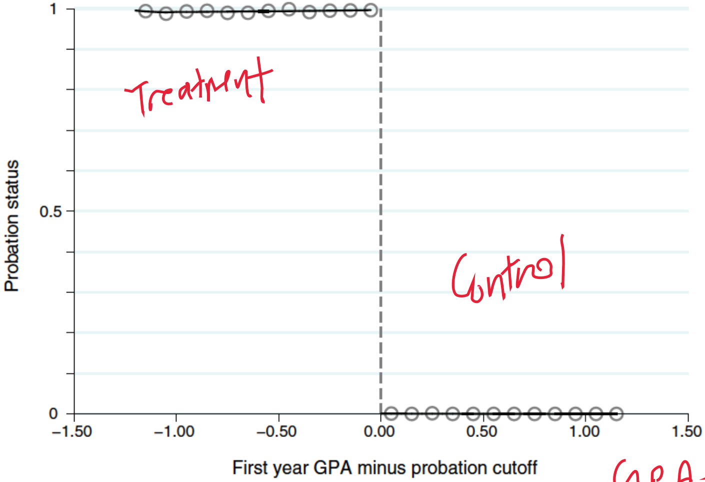
Treatment : $GPA \in [1.2, 1.5)$

Control : $GPA \in [1.5, 1.8]$

$$\bar{Y}_{treat} - \bar{Y}_{control}$$

Academic Probation and First Year CGPA

Lindo et al. (2010) uses U of T data:



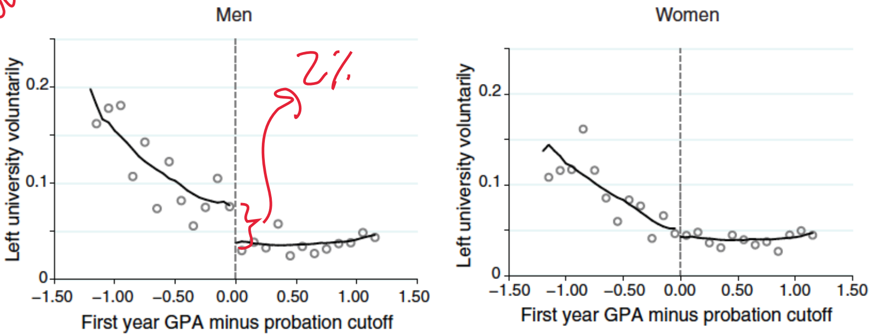
1.5
=
GPA − C

Academic Probation and Dropout

↑GPA ⇒ ↓Dropout

Effects of academic probation on leaving university by gender

Dropout



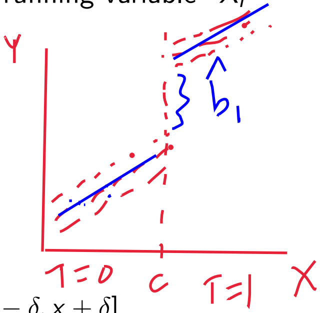
Probation increases dropout rate by 2% for men avg.

Y = College grad, T = HS scholarship, X = Score

Regression Discontinuity Design

- Treatment assignment T_i depends on "running variable" X_i

$T_i \in \{0,1\}$



- $T_i = T_i(X_i) = I(X_i \geq c)$
    - Treatment is discontinuous at $X_i = c$
- $Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \epsilon_i$
    - Can constrain data such that $X_i \in [x - \delta, x + \delta]$

$\hat{Y}_{i,T=0} = \hat{b}_0 + \hat{b}_2 X_i$, $\hat{Y}_{i,T=1} = \hat{b}_0 + \hat{b}_1 + \hat{b}_2 X$

- OLS estimate $\widehat{b}_1$ has causal interpretation
    - Estimates ATE at cutoff $X_i = c$