Causal Inferece

M. Fuat Kına

Problem of causality

- What is the difference between observational data and experimental designs?
  - Potential outcomes framework
  - How to know about counterfactuals

- We have different observations whether they are exposed to the treatment (treated) or not (untreated).

- Experiments depends on two potential outcomes for the same unit, as being treated and untreated, and the difference would give the treatment effect.

Treatment effect

- Consider the direct effect of a program on potential participant “i”
- Let’s think about the outcome variable Y for the participant i in two states of the universe
  - Y(0)i if i does not participate in the program
  - Y(1)i if i participates in the program

- Y(1)i – Y(0)i is the effect of the treatment on i
- The effect of the treatment on i is i‘s outcome if i is treated, compared to i‘s outcome if i is not treated

Potential outcomes

- The potential participant i either participates in the program or not
- One of the two potential outcomes is observed but the other one is not
  - If i is a participant, Y(1)i is observed and Y(0)i is not
  - If i is a nonparticipant, Y(0)i is observed and Y(1)i is not

- The unobserved outcome is called the counterfactual outcome

Counterfactuals are unobservable

- Y(1)i − Y(0)i is never observed

  We observe whether

  - Y(1)i or
  - Y(0)i
- But never both, since the counterfactual is unobservable. As a result, the effect of any program/intervention on a potential participant is fundamentally unobservable.

- Most of regression models have endogeneity problem
  - omitted variable bias
  - reverse causality

Natural experiments

- Since it is almost impossible to mention on causality in conventional models, scholars frequently use these words: prediction, explanatory power, association, etc. However, researchers still need to argue about causality; therefore, to provide an experimental framework.
- There are some techniques to construct experimental designs by observational data. They are not experimental, but semi- or quasi-experimental. Also called as **natural experiments**.
- Natural experiments help to make assumptions about counterfactuals.
- Requires an exogenous well-fitting natural designs.

Four quasi-experimental techniques

- *Matching*
- *Instrumental variable*
- *Regression discontunity design*
- *Difference in differences*

1. Matching

The simplest technique

Only for dummy treatments

Requires too many control variables

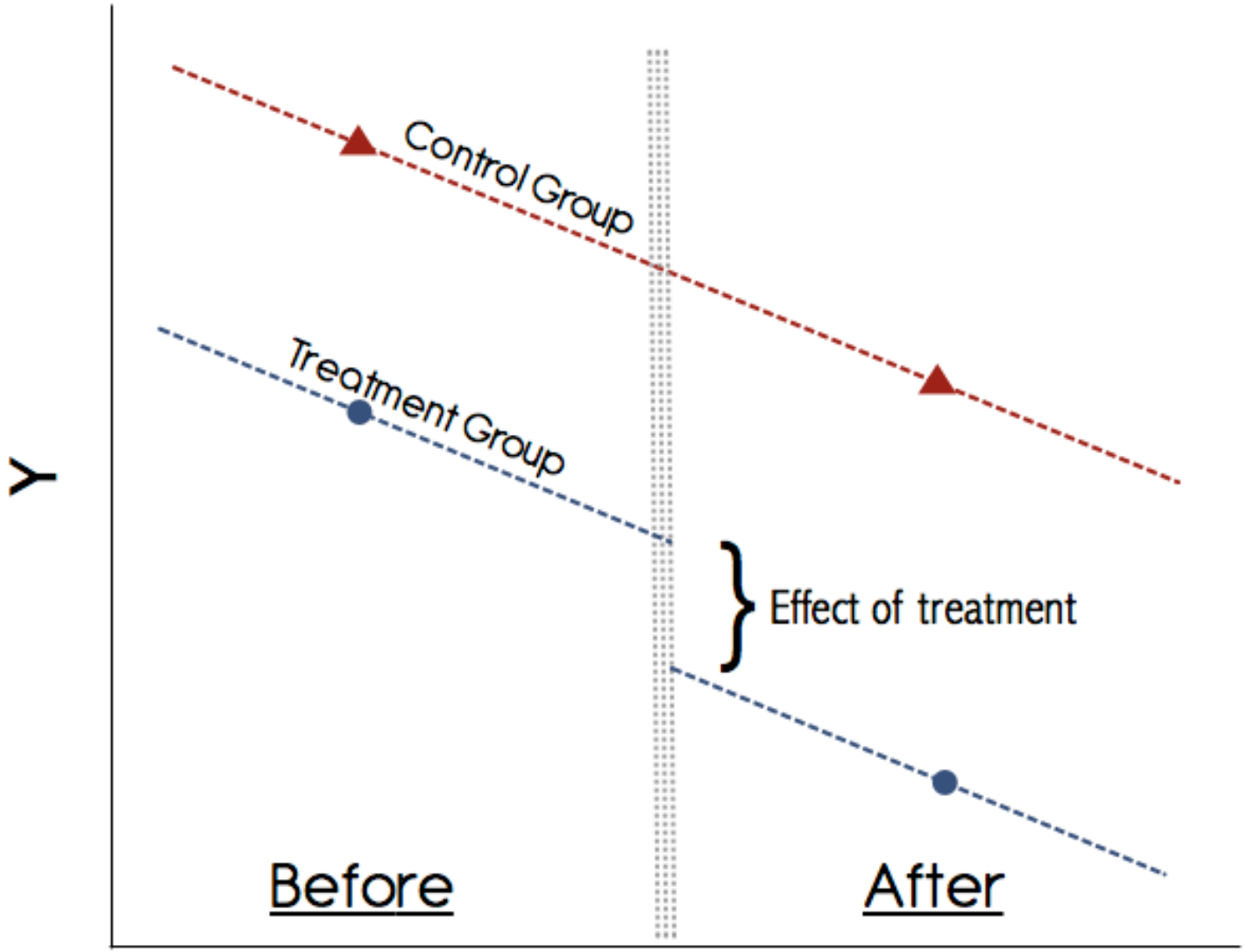
2. Difference in differences

	Pre-Treatment	Post-Treatment
Treatment Group	A (not yet treated)	**B (treated!)**
Control Group	C (never treated)	D (never treated)

- Needs at least two time periods, one for post- and one for pre-treatment.
- Needs at least two units, one would be treated and another one would be untreated.

2. Difference in differences

- The impact of Syrian refugees on unemployment
- The recession and inflation violates our findings
- Compare similar cities closed to the border and not (Antep and Erzurum)
- Closeness to the border is an exogenous dimension



3. Regression discontunity design

- Needs a natural treshold
- Compares cases located just the above of treshold with other cases just at the below of it.
- Let's assume we want to measure the impact of high-school education on grades in the university entrance exam.
- But we do not know anything about variation in students' ability, going to a high school is a totally self-selected process.

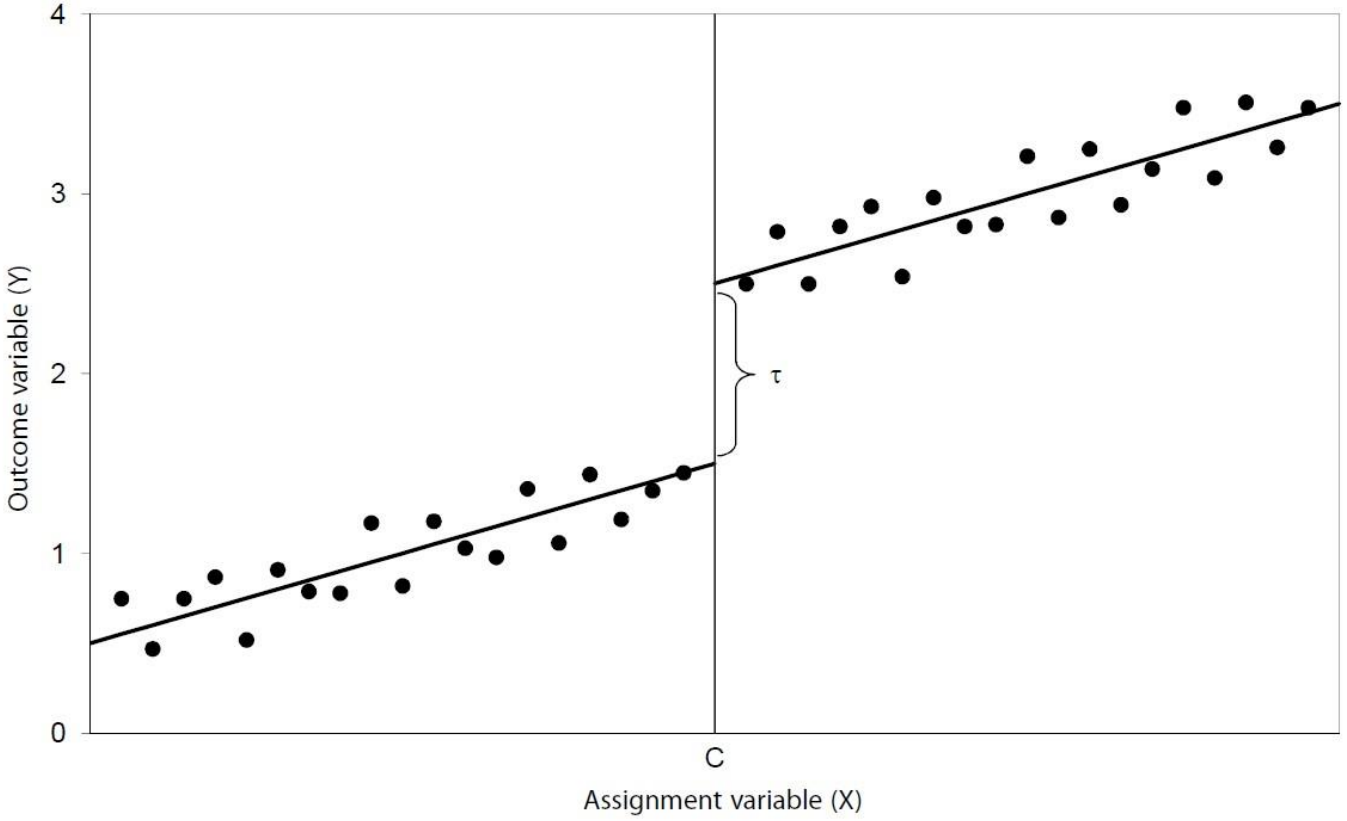


Figure 14.1 Simple linear RD setup

4. Instrumental variable (IV-2SLS)

- Requires a variable that correlates with independent variable, and uncorrelated with dependent variable
- It is so hard to justify, always open to criticism

- Let’s assume we are running an experiment. Even though they are randomly assigned to the treatment, some guys do not take the treatment. And some others do the opposite. The exogenous instrumental variable should exactly act as this randomized assignment.

Post-estimation

- There is no powerful measure to test statistical validity of quasi-experimental designs.
  - Always there are some, but they provide insufficient evidence to believe.
- Natural experimental designs are totally about the narrative.
  - How to be justified, to what extent the framework seems like an experiment?
- Scholars always have to deal with causality in their statistical analyses, after or before they present their regression results.

Use-case: Instrumental variable (IV-2SLS)

- Explicitly provide case-specific reasons for using a natural experiment
- Construct a historical/theoretical/social narrative for your design
- Exclusion restriction
  - Convince readers that you have an “exogenous” dimension
  - Why your instrument is uncorrelated with the DV, and why it is strong predictor of the IV?

What is the impact of MST's land invasions on the Bolsa Familia social assistance program?

- MST: A radical social movement in rural Brazil, invading lands of big landowners in the name of landless poor since the mid-1980s.
- Land reform: What the MST claims. Government buys the land from the landowners and recognizes the use-right of invaders.
- Bolsa Familia: One of the largest social assistance programs in the world, implemented in 2004.
- Argument: The Brazilian government have used Bolsa Familia program to demobilize (pacify or contain) the rural unrest, as a substitute for the MST's demand for land reform.
- What can be a proper instrument?

Let’s run an IV2SLS model in python

Yörük, Gürel, Kına (underreview at AJS). Demobilization by Substitution.

- First stage / AITTE: $X_{1,i,t-1} = \pi_0 + \pi_1 Z_{i,1995} + \alpha_n C_{n\ i,t-1} + \eta_{1,k} + \eta_{2,t} + \upsilon_{i,t}$
- Second stage / LATE: $Y_{i,t} = \beta_0 + \beta_1 X_{1,i,t-1} + \lambda_n C_{n,i,t-1} + \eta_{1,k} + \eta_{2,t} + u_{i,t}$

- $Y_{i,t}$ is Bolsa Família in a given i^{th} municipality and t^{th} year,
- $X_{1,i,t-1}$ is the number of cumulative land invasions in a given i^{th} municipality and $t\text{-}1^{th}$ year,
- β_1 is the causal effect of land invasions on the dependent variable,
- $C_{n,i,\ t-1}$'s are n-number control variables for a given i^{th} municipality and $t\text{-}1^{th}$ year,
- Z_i is the instrumental variable: the size of intended lands in 1995 (hectare) in a given i^{th} municipality,
- π_1 is the effect of the instrumental variable on land invasions in the first stage of 2SLS,
- π_0 and β_0 are the constants of the first and second stages of 2SLS, respectively,
- $\upsilon_{i,t}$ and $u_{i,t}$ are the error terms of the first and second stages of 2SLS, respectively, $\eta_{1,k}$ is year-invariant, state-specific factors, and $\eta_{2,t}$ is municipality-invariant, year-specific factors.