Political Science 209 - Fall 2018

Observational Studies

Florian Hollenbach

17th September 2018

What is the fundamental problem of causal inference?

Review

What about randomized control trials allows us to credibly estimate a causal effect?

Get out the Vote Study

What can induce citizens to vote?

What was the experiment?

What was the experiment?

Letters to randomized households with treatment:

1. Naming and Shaming: your neighbors will know
2. Civic Duty
3. Hawthorne Effect Message
4. Control (no letter)

Let’s go to R-studio quick

What is the main problem for observational studies?

What is the main problem for observational studies?

- Confounders: variables that are associated with both treatment and outcome

What is the Problem with Confounders?

What is the Problem with Confounders?

- If pre-treatment characteristics are associated with treatment and outcome, we can't disentangle causal effect from confounding bias

What is the Problem with Confounders?

- If pre-treatment characteristics are associated with treatment and outcome, we can't disentangle causal effect from confounding bias
- Selection into treament example: Maybe minimum wage was increased because unemployment was particularly low in NJ, but not PA

Examples of Confounding

- Are incumbents more likely to win elections? Yes, but...

Examples of Confounding

- Are incumbents more likely to win elections? Yes, but...

- Incumbents receive more campaign contributions
- Incumbents have more staff

Examples of Confounding

- Does higher income lead countries to democratize?

Examples of Confounding

- Does higher income lead countries to democratize?
- Higher income countries have more educated populations

What can we do about confounding in observational studies?

What can we do about confounding in observational studies?

- Make *Treatment* and *Control* groups as similar to each other as possible
- Especially on variables that might matter for treatment status and outcome
- Analyze subsets or *statistical control*, such that we compare treated and control units that have same value on confounder

Another problem with observational studies:

- Reverse causality

Another problem with observational studies:

- Reverse causality
- Example: Does economic growth cause democratization or democratization cause growth?

Why do experiments not suffer from the threat of reverse causality?

Observational studies

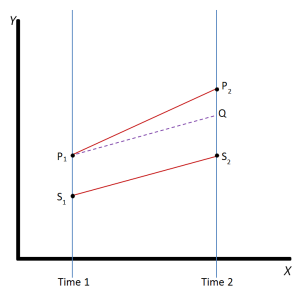
Difference-in-Differences Design

Difference-in-Differences Design

- Compare trends before and after the treatment across the same units
- Takes initial conditions into account

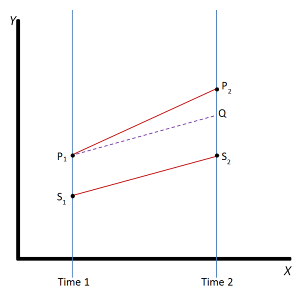
Difference-in-Differences Design

- Need data measured for both treatment and control at two different time periods: before and after treatment



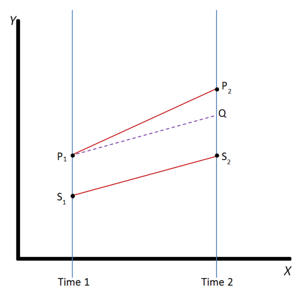
- Total difference between P2 and S2 can not be attributed to treatment. Why?

Difference-in-Differences Design



What might be a necessary condition for Diff-in-Diff to work?

Difference-in-Differences Design



What might be a necessary condition for Diff-in-Diff to work?

Parralel Trends Assumptions

Difference-in-Differences Design

The **difference-in-differences** (DiD) design uses the following estimate of the average treatment effect for the treated (ATT),

$$\text{DiD estimate} = \underbrace{\left(\overline{Y}^{\text{after}}_{\text{treated}} - \overline{Y}^{\text{before}}_{\text{treated}}\right)}_{\text{difference for the treatment group}} - \underbrace{\left(\overline{Y}^{\text{after}}_{\text{control}} - \overline{Y}^{\text{before}}_{\text{control}}\right)}_{\text{difference for the control group}}$$

The assumption is that the counterfactual outcome for the treatment group has a time trend parallel to that of the control group.

Describing numeric variables:

- Mean
- Median
- Quantiles

Quantiles

- splitting observations into equaly size groups, e.g., quartiles, quantiles
- 75th percentile is the threshold under which 75% of observations lie
- What percentile is the median?

Describing the spread of numeric variables:

- IQR:

Describing the spread of numeric variables:

- IQR:

Difference between 75th percentile and 25th percentile

Describing the spread of numeric variables:

Standard Deviation

Describing the spread of numeric variables:

Standard Deviation

$$\text{SD} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Standard Deviation

The sample **standard deviation** measures the average deviation from the mean and is defined as,

$$\text{standard deviation} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{or} \quad \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\bar{x}$ represents the sample mean, i.e., $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and n is the sample size. Few data points lie outside of 2 or 3 standard deviations away from the mean. The square of standard deviation is called **variance**.