Lecture 13

Alex Stephenson

10-10-2021

What did we cover last time?

For the last two weeks we have covered experiments with non-compliance

We now have three theoretical estimands: ATE, ITT, CACE.

We can estimate all three with data

What are we doing today?

Today we are going to review what we have covered over the semester so far

We are going to start with Week 1 and move to where we are at now

Framework for a Causal Question

"What is the effect of X on Y?"

This question needs to be interesting.

This question needs to be answerable in the real world

The way we answer it is a causal research design.

- A statement of how a study will estimate a relationship between two variables that is causal in nature.
- "How would I do this if it was possible run the ideal experiment?"

What do we need for a causal question?

1. What is the estimand I care about? ATE? ITT? CACE? Something else?

2. What is the target population of interest?

- For whom does this study generalize?

1. Why is learning this quantity useful?

- The so what question. There are lots of questions in the world, why should you spend time learning the answer to this one?

Causality and Manipulation

We do experiments because we follow the dictum that there is "no causation without manipulation"

Causality is an action applied to a unit.

We represent these actions with Potential Outcomes

$$Y_i(1)D + Y_i(0)(1 - D)$$

ATE

The main causal estimand we consider is the ATE. The sample analog is the SATE.

The ATE is $E[Y_i(1) - Y_i(0)]$ and has a causal interpretation when our assignment mechanism D is unconfounded with the potential outcomes.

In addition, we need to assume that it is the case that potential outcomes depend solely on whether a unit itself receives a singular treatment.

- Exclusion Restriction about treatment $Y_i(z_i = 1, d_i) = Y_i(z_i = 0, d_i)$
- Non-interference between units $Y_i(\mathbf{d}) = Y_i(d)$
- Know what the treatment is and guarantee it is the same for all units

Randomization

Randomization is a mechanism that ensures that either:

$$P[D_i = 1] = p, \forall i$$
$$P[D_i = 1 | X = x] = f(x), \forall i$$

We either know this for sure, or we assert via the Conditional Independence Assumption that

$$Y_i(0), Y_i(1) \perp!!! \perp D_i | \mathbf{X}_i, \forall i$$

Analyze as you randomize as you design

We start with the question, which leads to reasons to prefer some random assignment rather than another.

Complete Random Assignment: a procedure that guarantees that exactly m of N total units are assigned to treatment with equal probability.

Block Random Assignment: A procedure where units are partitioned into sub-groups (blocks) and complete random assignment occurs within each block.

Cluster Random Assignment: Assign units to clusters. All units in the same cluster are placed as a group into either the treatment or control conditions.

Randomization Inference

Because our designs are estimates on a sample to learn something about a population, we need to consider inference.

Randomization inference is a procedure that can be used to test the sharp null hypothesis that for all units, the ITE is exactly 0.

For randomization inference, we compute the sampling distribution of a test statistic of interest. The probability of obtaining a test statistic at least as large as the observed test is the p-value.

Conventionally, if the p-value is below a certain level we reject the null hypothesis.

Regression

Regression is a device that allows us to estimate Average Treatment Effects

We started with the equivalence between a Difference in Means and regression in the binary case.

We then noted that the way to estimate block fixed effects is a regression of the form $Y \sim D + X_1 + X_2 + \ldots + X_n$

Regression

We saw that we can motivate regression in several different ways:

1. The best linear estimate for the Conditional Expectation Function
2. Regression as a natural way to estimate an experiment
3. Regression as an information problem

Regression and Interactions

Sometimes our treatment depends on an interaction. An effect might be an interaction between two variables.

Regression can model this by simply adding another variable that is the product of the variables of interest.

We always need to estimate the main effect of each variable separately if we plan to include their interaction.

Regression and Omitted Variables

When we cannot guarantee unconfoundedness between treatment and potential outcomes, we need to worry about omitting a relevant variable.

Omitted Variable Bias is the difference between the true value and our estimate when we do not include a relevant variable.

Bad Controls

A bad control is a variable that is itself an downstream effect of the treatment variable. Bad controls are functionally identical to mediator variables.

Including a bad control in a regression leads to bias, and sometimes substantial bias.

Compliance

In an ideal experiment every unit follows directions. In real experiments, that rarely happens.

We have two kinds of non-compliance:

- One sided non-compliance: Some units in the treatment group do not receive treatment
- Two sided non-compliance: Some units in both treatment and control do not receive their intended treatment.

Compliance

In general we have four kinds of units:

Compliers: Units that always take the treatment they are assigned.

Never Takers: Units that never take treatment but will always take control.

Always Takers: Units that always take treatment and will never take control.

Defiers: Units that do the opposite of their treatment assignment.

ITT

When we run an experiment, we can also ask "Does the random assignment produce an effect?"

ITT: The causal effect of treatment assignment.

CACE

The average treatment effect for the subset who units who are compliers.

In general, if we assume monotonicity we can recover the CACE.

When the share of compliers is close to 0, slight violations of the exclusion restriction will lead to massive bias in our estimates. This is a major problem in observational research, and a reason to be skeptical of experiments with low compliance.

Estimating the CACE

We can estimate the CACE by hand as a ratio estimator:

$$CACE = \frac{ITT}{ITT_D}$$

We can also estimate this in R with the estimatr package

```
iv_robust(Y ~ D | Z, data = data)
```