Section 1: The CIA, Causality, Matching

Valentine Gilbert

February 3, 2022

Overview

Potential outcomes

Potential outcome notation is the foundation of everything we'll do this semester

- Suppose we observe outcome Y_i and treatment status $D_i \in \{0, 1\}$ for individual i
- Potential outcomes notation gives us a way of connecting **causal effects** for individual i to observed data

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

- Another way of writing this:

$$Y_i = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{\text{treatment effect}} D_i$$

- D_i doesn't have to be binary – we could have Y_{0i}, Y_{1i}, Y_{2i}...
- Alternative notation: $Y_i(0)$, $Y_i(1)$...

Estimands, estimators, and estimates

An **estimand** is something we want to estimate

- It often depends on things we can't observe
- E.g. the $ATE = E[Y_{1i} - Y_{0i}]$ is an estimand

An **estimator** is a way of estimating the estimand with observables

- It's a function that maps the data to an **estimate**
- Whether it's a consistent and/or unbiased estimator depends on assumptions
- E.g. $\widehat{ATE} = \frac{1}{N_1}\sum_{D_i=1} Y_i - \frac{1}{N_0}\sum_{D_i=0} Y_i$ is an estimator of the ATE
  - It's a consistent and unbiased estimator if D_i is independent of potential outcomes
  - $\widehat{ATE} \xrightarrow{p} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i} - Y_{0i}]$
  - $E[\widehat{ATE}] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i} - Y_{0i}]$

Causal estimands

- Any one individual's treatment effect $Y_{1i} - Y_{0i}$ is *fundamentally unknowable*
- But we *can* identify different average treatment effects
- In class we've talked about the average treatment effect (ATE), the average treatment effect on the treated (TOT/ATT), and the average treatment effect on the untreated (TOU/ATU)
- Other treatment effects of interest include:
  - Conditional Average Treatment Effect (CATE)
  - Intent to treat (ITT)
  - Local average treatment effect (LATE)
- Which kind of treatment effect is most relevant depends on the policy change we're considering

ATE and CATE

- The ATE is the average individual treatment effect in the population of interest:

$$ATE \equiv E[Y_{1i} - Y_{0i}] = E[TE_i]$$

- E.g., the average effect of college attendance on earnings for the population
- This is generally what we estimate in an RCT
- The CATE is average individual treatment effect *conditional* on some characteristic:

$$CATE \equiv E[TE_i | X_i = x]$$

- E.g., the treatment effect of college attendance on the earnings of low-income students
- We can estimate this in an RCT by comparing treatment and control group means for different groups in the study

TOT and TOU

- The TOT is the average individual treatment effect conditional on being treated:

$$TOT \equiv E[TE_i | D_i = 1]$$

- E.g., the effect of college attendance on the earnings of those who attend college
- In an RCT, the TOT is equal to the ATE
- The TOU is the average individual treatment effect conditional on being untreated:

$$TOU \equiv E[TE_i | D_i = 0]$$

- E.g., the effect of college attendance on the earnings of those who don't attend college
- Again, randomization of treatment means that the TOU is equal to the ATE

ITT and LATE

- Often, we can't ensure perfect compliance with assigned treatment status. The ITT is the average effect of being *offered* treatment:

$$ITT \equiv E[Y_i(\text{Offered treatment}) - Y_i(\text{Not offered})]$$

- E.g., the effect of college *admission* on earnings
- The LATE is the average individual treatment effect among individuals "local" to the margin of treatment:

$$LATE \equiv E[Y_{1i} - Y_{0i} | complier]$$

- E.g., the effect of college attendance among those who attended *because* they won a scholarship
- The ITT and LATE are important causal estimands for RCTs with imperfect compliance and for instrumental variables methods

The CIA and causality

The Conditional Independence Assumption (CIA) says that treatment is uninformative about potential outcomes, conditional on some set of observed characteristics

- Informally, treatment is "as good as randomly assigned" among similar individuals
- In math: $\{Y_{1i}, Y_{0i}\} \perp D_i | X_i$
- The CIA is crucial to interpret *any* estimate causally – not just in observational studies
- Other identifying assumptions (e.g. "parallel trends" for DID, continuity for RDD) are assumptions needed for CIA to hold in those contexts

Interpreting Estimates

- It's important not to use causal language when an estimate isn't causal
- But you **should** use causal language when interpreting causal estimates
- Causal language:
  - "college attendance *increases* earnings"
  - "college attendance *leads to*..."
  - "The *causal effect* of college attendance is..."
- Descriptive language:
  - "college attendance *is associated with*..."
  - "college attendees *earn more than* non-attendees"
- When using causal language:
  - **Do** describe who the estimates apply to
  - **Don't** say "holding X constant", "controlling for X", or "conditional on X"
- When using descriptive language:
  - **Do** say "holding X constant", etc.
  - **Don't** say "assuming conditional independence holds", etc.

Deriving a matching estimator: Overview

Suppose our estimand of interest is the effect of treatment on the treated:

$$\delta_{TOT} \equiv E[Y_{1i} - Y_{0i} | D_i = 1]$$

- To derive an estimator, we need to write this in terms of observable quantities.
- Assuming $\{Y_{0i}, Y_{1i}\} \perp D_i | X_i$ allows us to rewrite δ_{TOT} as the expectation of conditional differences in means.
- When X_i is discrete, the estimand is a weighted average of conditional differences in means.

Deriving a matching estimator: Derivation

- We begin by iterating expectations:

$$\begin{aligned}\delta_{TOT} &\equiv E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E\{E[Y_{1i} - Y_{0i}|X_i, D_i = 1]|D_i = 1\} \\ &= E\{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 1]|D_i = 1\}\end{aligned}$$

- What do I know about $E[Y_{0i}|X_i, D_i = 1]$?

$$\begin{aligned}\delta_{TOT} &= E\{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 0]|D_i = 1\} \\ &= E\{\underbrace{E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0]}_{\delta_X}|D_i = 1\}\end{aligned}$$

- If X_i is discrete, then:

$$\delta_{TOT} = \sum_x \delta_x P(X_i = x|D_i = 1) = \frac{\sum_x \delta_x P(D_i = 1|X_i = x)P(X_i = x)}{\sum_x P(D_i = 1|X_i = x)P(X_i = x)}$$

- Do you anticipate any problems when trying to estimate δ_{TOT}?

Regression as matching estimator: Overview

- How does our regression estimand relate to the estimand of interest, δ_{TOT}?
- We'll see that the regression estimand can also be written as a weighted average of conditional mean comparisons.
- The weights in this weighted average are proportional to the variance of treatment within each covariate cell.

Regression as matching estimator: Derivation

- Let's assume that our model is saturated in X_i, so that $E[D_i|X_i] = X_i\beta$
- We start with the regression anatomy formula:

$$\begin{aligned}
\delta_R &= \frac{cov(Y_i, \tilde{D}_i)}{var(\tilde{D}_i)} \\
&= \frac{E[Y_i\tilde{D}_i]}{E[\tilde{D}_i^2]} \\
&= \frac{E\{(D_i - E[D_i|X_i])Y_i\}}{E\{(D_i - E[D_i|X_i])^2\}} \\
&= \frac{E\{(D_i - E[D_i|X_i])E[Y_i|D_i, X_i]\}}{E\{(D_i - E[D_i|X_i])^2\}}
\end{aligned}$$

Regression as matching estimator: Derivation

- Let's focus on the numerator of the previous expression, $E\{(D_i - E[D_i|X_i])E[Y_i|D_i, X_i]\}$
- We can expand the CEF like so: $E[Y_i|D_i, X_i] = E[Y_i|D_i = 0, X_i] + \delta_X D_i$
- Substituting this into the numerator gives us:

$$
\begin{aligned}
&E\{(D_i - E[D_i|X_i])E[Y_i|D_i, X_i]\} \\
&\quad = E\{(D_i - E[D_i|X_i])E[Y_i|D_i = 0, X_i]\} \\
&\qquad + E\{(D_i - E[D_i|X_i])D_i\delta_X\}
\end{aligned}
$$

- The first term simplifies to 0. Why?
- The second term simplifies to $E[(D_i - E[D_i|X_i])^2\delta_X]$:

$$
\begin{aligned}
E\{(D_i - E[D_i|X_i])D_i\delta_X\} &= E\{(D_i - E[D_i|X_i])^2\delta_X\} \\
&\quad + E\{(D_i - E[D_i|X_i])E[D_i|X_i]\delta_X\}
\end{aligned}
$$

Regression as matching estimator: Derivation

- Having simplified the numerator, the regression estimand looks like this:

$$\delta_R = \frac{E[(D_i - E[D_i|X_i])^2 \delta_X]}{E[(D_i - E[D_i|X_i])^2]}$$
$$= \frac{E[E[(D_i - E[D_i|X_i])^2|X_i]\delta_X]}{E[E[(D_i - E[D_i|X_i])^2|X_i]]} = \frac{E[\sigma_D^2(X_i)\delta_X]}{E[\sigma_D^2(X_i)]}$$

- Like with the matching estimator, we can express the estimand when X_i is discrete as a weighted average:

$$\delta_R = \frac{\sum_x \delta_x \sigma_D^2(x) P(X_i = x)}{\sum_x \sigma_D^2(x) P(X_i = x)}$$

- Do you anticipate any problems with trying to estimate δ_R?

Comparing δ_{TOT} and δ_R

Let's compare our two estimands side by side:

$$\delta_R = \frac{\sum_x \delta_x \sigma_D^2(x) P(X_i = x)}{\sum_x \sigma_D^2(x) P(X_i = x)} \qquad \delta_{TOT} = \frac{\sum_x \delta_x P(D_i = 1 | X_i = x) P(X_i = x)}{\sum_x P(D_i = 1 | X_i = x) P(X_i = x)}$$

- How are these two estimands similar and how are they different?
- Which covariate cells get higher weights in the regression estimand relative to the TOT estimand?
- What are the advantages and disadvantages of each weighting scheme?

Average Weighter
@agoodmanbacon

Every day I stray farther from the view that controlling for things using a regression is a good idea.

2:42 PM · Feb 1, 2022 · Twitter for iPhone

Amanda Stevenson
@ajeanstevenson

Many things got easier once I accepted that regression is just fancy averaging.

2:53 PM · Feb 1, 2022 · Twitter Web App

Aaron Schwartz
@A_Schwa

Econ twitter: "Here's a 100-page working paper full of dazzling creativity and sophistication"

Also econ twitter: "Any day now, we'll figure out what the heck a regression does."

4:13 PM · Feb 1, 2022 · Twitter Web App