Harris School of Public Policy

Program Evaluation

Jaskirat Kaur

TA session 2
Oct 26, 2023

Outline

- Review of Fundamental Problem of Causal Inference
- Average Treatment Effect (ATE) vs Naïve Estimator
- ATT and ATN
- Omitted Variable Bias (OVB)
- R Markdown Tips

Fundamental Problem of Causal Inference

- The true treatment effect T_i for person i would be the difference in their outcomes Y between when they are treated (1) and not treated (0):

$$T_i = Y_1 - Y_0$$

- We cannot observe both Y_1 and Y_0.
- We ONLY observe Y_1 OR Y_0, not both.

Average Treatment Effect (ATE) vs Naïve Estimator

- The Average Treatment Effect (ATE):

$$T^{\mathrm{ATE}} = \mathbb{E}[Y_1 - Y_0]$$

- Naïve Estimator of ATE:

$$T_{\text{naïve}} = Y_1 - Y_0$$

Average Treatment Effect (ATE) vs Naïve Estimator

- The ATE measures potential outcomes, while the naïve uses observed outcomes.
- We can think of the naïve estimator as:

$$T_{\text{naïve}} = T^{\text{ATE}} + \text{Selection Bias}$$

Selection on Observables vs Unobservables

- Selection on observables: when the "selection bias" happens along a set of variables we can observe
- We can minimize the difference between our unattainable T^{ATE} and our attainable T_{nave} by controlling for these observable variables
- Selection on unobservables: when the "selection bias" happens along a set of variables we cannot observe
- Example: motivation (hard to observe and accurately quantify/measure)
- Controlling for variables alone will not fix the issue

ATT and ATN

- Let's consider a sample of 100 people: 60 who attended college, 40 who did not
- Treated: 60 who attended college
- Untreated: 40 who did not attend college
- Average Treatment Effect on Treated (ATT): the T^{ATE} for the 60 people who attended college
- Average Treatment Effect on Not-Treated (ATN): the T^{ATE} for the 40 people who did not attend college

ATT and ATN

- Average Treatment Effect on Treated (ATT): the T^{ATE} for the 60 people who attended college
- Average Treatment Effect on Not-Treated (ATN): the T^{ATE} for the 40 people who did not attend college
- How would these relate to the ATE?

ATT and ATN

- Average Treatment Effect on Treated (ATT): the T^{ATE} for the 60 people who attended college
- Average Treatment Effect on Not-Treated (ATN): the T^{ATE} for the 40 people who did not attend college
- How would these relate to the ATE?
- The ATE would simply be a weighted average of the ATT and the ATN - we can think of the average treatment effect of our 100 people as a combination of the ATE for people who were treated, and ATE for people who weren't
- $\text{ATE} = 0.6(\text{ATT}) + 0.4(\text{ATN})$

ATT and ATN

- The following table can help illustrate this point:

	Attended College($D = 1$)	Did Not Attend College($D = 0$)
$Y_i(0)$	Can't Observe	40,000
$Y_i(1)$	80,000	Can't Observe

Table: ATT versus ATN

ATT and ATN

- The following table can help illustrate this point:

	Attended College($D = 1$)	Did Not Attend College($D = 0$
$Y_i(0)$	Can't Observe	40,000
$Y_i(1)$	80,000	Can't Observe

Table: ATT versus ATN

- Suppose we figured out some research design that would help us estimate the counterfactuals that we cannot observe
- What is the ATT? The ATN?

ATT and ATN

- Let's say our research design yielded the following estimates:

	Attended College($D = 1$)	Did Not Attend College($D = 0$)
$Y;(0)$	60,000	40,000
$Y;(1)$	80,000	30,000

Table: ATT versus ATN

- What is the ATT? The ATN? - The ATT is 20,000, but the ATN is -10,000

Omitted Variable Bias

- Omitted Variable Bias (OVB) occurs when a relevant variable is left out of the model.

Omitted Variable Bias

- Suppose the true relationship between variables is defined by the model: $Y = b_0 + b_1 X_1 + b_2 X_2 + e$
- Maybe it's impossible to get data for X_2, so we instead run: $Y = a_0 + a_1 X_1 + e$

Omitted Variable Bias

- Suppose the true relationship between variables is defined by the model: $Y = b_0 + b_1 X_1 + b_2 X_2 + e$
- Maybe it's impossible to get data for X_2, so we instead run: $Y = a_0 + a_1 X_1 + e$
- Intuition of omitted variable bias: how far off a_1 is from b_1 in the regression model.

Omitted Variable Bias

- Consider the example:

$$\text{salary} = b_0 + b_1(\text{college attendance}) + b_2(\text{motivation}) + e$$

- We cannot measure "motivation", so we instead run:

$$\text{salary} = a_0 + a_1(\text{college attendance}) + e$$

Omitted Variable Bias

- The bias of our estimate for b_1 will be the product of the relationship between motivation and salary, and motivation and college attendance.
- Likely positive bias.

Fundamental Problem of Causal Inference

- As you have heard several times by now: the biggest problem around causal inference is that we only observe ONE state of the world
- Let's consider an example:
  - Person i (Bob)
  - Treatment D_i (training program)
  - Outcome Y_i (income level)

State of the World 1: Bob attends training program

- In words:
  - Observed outcome: income level when Bob attends training program
  - Unobserved outcome: income level when Bob doesn't attend training program
- In math:
  - Observed Outcome $= Y_i(D_i = 1)$
  - Unobserved Outcome $= Y_i(D_i = 0)$

State of the World 2: Bob DOES NOT attend training program

- In words:
  - Observed outcome: income level when Bob doesn't attend training program
  - Unobserved outcome: income level when Bob attends training program
- In math:
  - Observed Outcome $= Y_i(D_i = 0)$
  - Unobserved Outcome $= Y_i(D_i = 1)$

What is Bob's treatment effect t_i?

- In words:
  - It's the difference in Bob's income level in the state of the world with the training program, and Bob's income level in the state of the world without the training program.
- In math:
  - $t_i = Y_i(D_i = 1) - Y_i(D_i = 0)$
- Since we cannot observe both states of the world at once, this is impossible to calculate.

Since individual TE are impossible $\rightarrow$ ATE

- As we just saw, we cannot calculate individual treatment effects.
- But, we can calculate average treatment effects across many individuals.
- $T_{\text{ATE}} = \mathbb{E}[Y_i(D_i = 1) - Y_i(D_i = 0)]$
- In order to get an unbiased estimate of the ATE, we need some conditions and assumptions.

Working with R/Rmd

- Let's move on to discussing R!