PPHA 34600 TA Session Week 1

Zhijie Yan (based on previous material)

4/9/2020

1. Fundamental Problem of Causal Inference

The fundamental problem of causal inference is missing counterfactuals, i.e., we do not observe the outcome of the same individual with **and** without the treatment.
Denote the treated outcome of individual i as $Y_i(1)$ and the untreated outcome of person i as $Y_i(0)$, the treatment effect on unit i is

$$\tau_i = Y_i(1) - Y_i(0)$$

However, we either observe $Y_i(1)$ or $Y_i(0)$. As a result, we can **never** find the treatment effect for an individual.

2. ATE, ATT, ATNT, and the Naïve Estimator

Though we do not observe τ_i, we do observe the marginal distribution of $Y(1)$ and $Y(0)$. That is, we observe the distribution of treated outcomes (from the treatment group), as well as the distribution of untreated outcomes (from the control group). However, we **have no idea about** the joint distribution of $Y(1)$ and $Y(0)$ - if we observe $Y_i(1)$, we do not see $Y_i(0)$, and vice versa.

Average Treatment Effect (ATE)

The treatment effect might vary across individuals, and ATE is simply the average of τ_i for **all individuals**, that is

$$\tau^{ATE} = E[Y_i(1) - Y_i(0)].$$

Again, the fundamental problem of causal inference! It would be nice if we compare individuals to themselves, then any change in the outcome will be due to the factor(s) that changed. **But we do not observe potential outcomes.**

The Naïve estimator

What can we do with what we observe? We calculate the naïve estimator, which is

$$\tau^N = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$$

where D_i is the treatment variable.
In other words, the naïve estimator is the difference between the average treated outcome on the treated and the untreated outcome on the untreated. **It does not involve any potential outcomes.** Isn't it nice? **Of course not!**
It is naïve because it implicitly assumes that people who received the treatment are on average the same as

those who did not, or in math

$$E[Y_i(1)] = E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0],$$

$$E[Y_i(0)] = E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0].$$

Most of the time, there is selection, which makes the treatment group and control group different. If it's selection on the observables, we may control for these variables. If it's selection on the unobservables, we need to do better than adding covariates, which you will learn throughout the course.

Average Treatment Effect on the Treated (ATT)

It is natural to think that the average treatment effects are different for different groups of people. Sometimes we are interested in the average treatment effect on the treated, which is

$$\tau^{ATT} = \underbrace{E[Y_i(1)|D_i = 1]}_{\text{observed}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{potential/unobserved}}$$

Unfortunately, τ^{ATT} involves potential outcomes.
Now I want to compare the naïve estimator and ATT,

$$\begin{aligned}
\tau^{ATT} &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] \\
&= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] \underbrace{+E[Y_i(0)|D_i = 0] - E[Y_i(0)|D_i = 0]}_{\text{adding and subtracting the same item - 0!}} \\
\text{rearranging to get} &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] + E[Y_i(0)|D_i = 0] - E[Y_i(0)|D_i = 1] \\
&= \tau^N + \underbrace{E[Y_i(0)|D_i = 0] - E[Y_i(0)|D_i = 1]}_{\text{selection bias}}
\end{aligned}$$

The naïve estimator is not an unbiased estimate of ATT because the average observed outcome of the control group is not the same as the average potential outcome of the treatment group.

Average Treatment Effect on the Non-Treated (ATNT)

We might also care about the average treatment effect on the non-treated, which is

$$\tau^{ATNT} = \underbrace{E[Y_i(1)|D_i = 0]}_{\text{potential/unobserved}} - \underbrace{E[Y_i(0)|D_i = 0]}_{\text{observed}}$$

Again, it involves potential outcomes. Similarly, we can compare the the naïve estimator and ATNT,

$$\begin{aligned}
\tau^{ATNT} &= E[Y_i(1)|D_i = 0] - E[Y_i(0)|D_i = 0] \\
&= E[Y_i(1)|D_i = 0] - E[Y_i(0)|D_i = 0] \underbrace{+E[Y_i(1)|D_i = 1] - E[Y_i(1)|D_i = 1]}_{\text{adding and subtracting the same item - 0!}} \\
\text{rearranging to get} &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] + E[Y_i(1)|D_i = 0] - E[Y_i(1)|D_i = 1] \\
&= \tau^N + \underbrace{E[Y_i(1)|D_i = 0] - E[Y_i(1)|D_i = 1]}_{\text{selection bias}}
\end{aligned}$$

The naïve estimator is not an unbiased estimate of ATT because the average potential outcome of the control group is not the same as the average observed outcome of the treatment group.

ATE, ATT, and ATNT

- Homogenous treatment effects: $\tau^{ATE} = \tau^{ATT} = \tau^{ATNT}$
- Heterogenous treatment effects: $\tau^{ATE} = Pr(D_i = 1)\tau^{ATT} + Pr(D_i = 0)\tau^{ATNT}$, i.e., the weighted average

Clear as Mud? Examples!

Many of you might have learned about the Roy Model when you were in stats class. Not sure if you still like it, but I'm using it as the example here. This time no economists or accountants, but a replication of the example from Lecture 2.

I will be generating some data here. The steps are as follows:

- Generating salaries for college attendees and non-attendees from a bivariate normal distribution (`MASS` package); same distributions as in lecture notes, $Y_i(1) \sim N(60000, 10000^2), Y_i(0) \sim N(65000, 5000^2), corr(Y_i(1), Y_i(0)) = 0.84$; 100000 samples.
- Let individuals "choose" between the incomes, if for person i, $Y_i(1) \geq Y_i(0)$, she will attend college, otherwise she does not. In this step, the package `tidyverse` is used to build a dataframe/tibble.
- Calculate the four estimators we discussed above and see how they are different from each other. In this step, I use the package `kableExtra` to make tables.

The results are summarized in Table 1.

Table 1: Average Observed and Potential Incomes

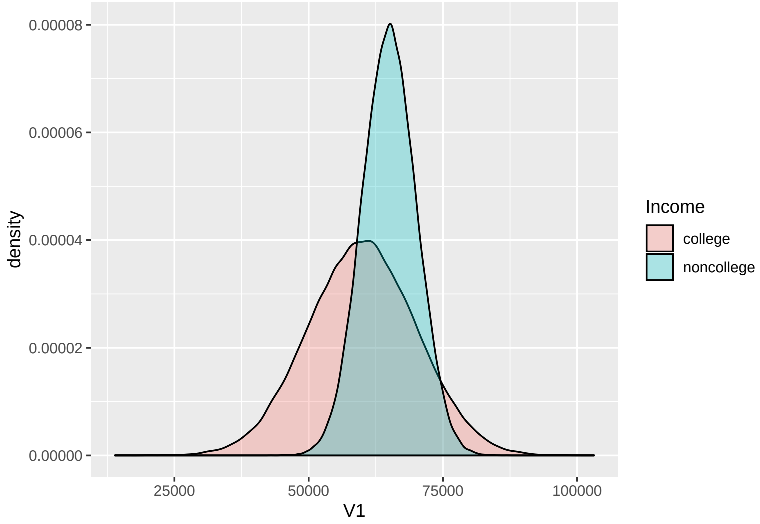
	Non-attendees	Attendees	Mean
College income	~~56600~~	72322	60014
Noncollege income	64000	~~68643~~	65008
Number of obs	78289	21711	100000

[1] Numbers in red are unobserved.
[2] All numbers are rounded to the nearest integer.

Now let's do some calculations (Use `` `r expression` `` for inline code so that you do not have to type the numbers yourself):

- $\tau^{ATE} = E[Y_i(1)] - E[Y_i(0)] = 60014 - 65008 = -4994$
- $\tau^{N} = E[Y_i(1)|attendees] - E[Y_i(0)|non\text{-}attendees] = 72322 - 64000 = 8322$
- $\tau^{ATT} = E[Y_i(1)|attendees] - E[Y_i(0)|attendees] = 72322 - 68643 = 3680$
- $\tau^{ATNT} = E[Y_i(1)|non\text{-}attendees] - E[Y_i(0)|non\text{-}attendees] = 56600 - 64000 = -7400$
- Verifying: $Pr(D = 1)\tau^{ATT} + Pr(D = 0)\tau^{ATT}$ = 0.21711 $\times 3680$ + 0.78289 $\times (-7400)$ = -4994 $= \tau^{ATE}$

These numbers vary a lot! Looking at the naïve estimator, we would force everyone into college. However, We already know that there is selection - people self sort into college because they know what is best for them. College attendees and non-attendees are different on average. They would have different average incomes had they all attended college or not. From the density plot, we can see that college attendees are on the two tails of the distribution. This is because the distribution of college attendees' incomes is more spread out.



3. OLS Recap

Before running regressions, let's review some ideas about OLS.

OLS Assumptions (Wooldridge, 2000)

- Linear in Parameters: the population model can be written as $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \varepsilon_i$
- Random Sampling: we have a random sample of observations $\{(X_{i1}, X_{i2}, ..., X_{ik}, Y_i) : i = 1, 2, ..., N\}$
- No Perfect Collinearity: In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.
- Zero Conditional mean: The error term has an expected value of zero given any values of the independent variables. In other words,

  $$E(\varepsilon_i | X_{i1}, ..., X_{ik}) = 0.$$

  These four assumptions give us consistent and unbiased estimates. ***(What does this mean?)***
- Homoskedasticity: The error term has the same variance given any value of the explanatory variables. In other words,

  $$Var(\varepsilon_i | X_{i1}, ..., X_{ik}) = \sigma^2.$$

The five assumptions are known as the Gauss-Markov assumptions. Under the five assumptions, OLS estimator $\hat{\beta}$ for β is the **best linear unbiased estimator (BLUE)**.

Simple Regression with a Dummy

The coefficients of a bivariate OLS can be obtained by minimizing the squared error

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_i)^2$$

In particular, when the independent variable is the treatment dummy, we have

$$\begin{aligned} E[Y_i|D_i = 1] &= \beta_0 + \beta_1 \times 1 + E(\varepsilon_i|D = 1), \\ E[Y_i|D_i = 0] &= \beta_0 + \beta_1 \times 0 + E(\varepsilon_i|D = 0), \\ \Rightarrow E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= \beta_1 + E(\varepsilon_i|D = 1) - E(\varepsilon_i|D = 0). \end{aligned}$$

The fourth assumption comes in handy, with $E(\varepsilon_i|D_i) = 0$, there will be no selection. This condition basically says that we should capture any variable that is correlated with the treatment and affects Y_i.
What if we didn't?

Omitted Variable Bias

Suppose that the true model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i,$$

But instead we omitted X_{i2} from our model and instead estimate

$$Y_i = \alpha_0 + \alpha_1 X_{i1} + u_i.$$

Then from the misspecified model, we have

$$\begin{aligned} E(\alpha_1) &= E[\frac{Cov(Y_i, X_{i1})}{Var(X_{i1})}] \\ &= E[\frac{Cov(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, X_{i1})}{Var(X_{i1})}] \\ &= \beta_1 + \beta_2 \frac{Cov(X_{i1}, X_{i2})}{Var(X_{i1})} + \underbrace{\frac{Cov(\varepsilon_i, X_{i1})}{Var(X_{i1})}}_{\text{Zero conditional mean}} \\ &= \beta_1 + \beta_2 \underbrace{\frac{Cov(X_{i1}, X_{i2})}{Var(X_{i1})}}_{\text{the slope by regressing } X_{i2} \text{ on } X_{i1}} \\ &\neq \beta_1 \text{ unless } \beta_2 Cov(X_{i1}, X_{i2}) = 0 \end{aligned}$$

Again, it tells us not to leave out variables that are correlated with our variable of interest **and** affect the outcome. For example, in the college example, suppose that working experience has a positive effect on income ($\beta_2 > 0$), and people who attend college will have less experience ($Cov(D_i, exper_i) < 0$). If we only regress incomes on college dummy, we would be understating the effect of attending college.

R and Regressions

With the simulated data we generated earlier, let's run a simple regression in R. In this section, the package `stargazer` will be used to obtain a nice regression table.

```r
# results = "asis": output as-is, i.e., write raw results from R into the output document
m <- lm(observed ~ D, data = df)
stargazer(m, title = "A Simple Model",
          dep.var.labels = "Income", # renaming the dependent variable
          header = F)                # get rid of the initial comments added by the author
```

Table 2: A Simple Model

	Dependent variable:
	Income
D	8,322.352***
	(38.552)
Constant	64,000.140***
	(17.963)
Observations	100,000
R^2	0.318
Adjusted R^2	0.318
Residual Std. Error	5,026.151 (df = 99998)
F Statistic	46,601.690*** (df = 1; 99998)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
# You can find a lot of stargazer tutorials, here's one that is pretty organized
# https://www.jakeruss.com/cheatsheets/stargazer/
```

Make sure that you know how to read the output!

4. More about R Markdown

There are many R Markdown references online. For example, RStudio provides a “Get Started” Tutorial here. You can transform a .Rmd file to other formats such as PDF, word, html by clicking **Knit**. Specifically, for a PDF document as the output, you need TeX distributions on your computer, which you can download here.

Now some basic syntax ~~which looks like nonsense~~. For example, the hashtags give us different levels of titles.

I am a big title because there is only one # before me!

Single star gives you italics, so does single underscore. Similarly, **double stars get you bold, so do double underscores.**

1. A numbered list, please put a blank line before you start
    - Now bullet points
        - a sublist
            - a sublist to the sublist
                - I can do it all day! (Actually don’t make it too deeply nested or you might get an error.)
        - You can use different symbols for the same level, as long as they are indented in the same way.
Two whitespaces will be enough. The return/enter key does not change the line.
You need two whitespaces after the last paragraph.
2. If you are familiar with LaTeX, it is the same to type some nice equations here
A equation not centered $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

    $$\text{A equation that is centered } y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

    Some symbols and more:

    $$\neq, \geq, \leq, \sim, \approx, \equiv, \pm, \rightarrow, \infty, \int_a^b, \sum_a^b, \prod_a^b, \frac{a}{b} \cdots$$

Now you can try to creat a R Markdown file of your own or play with this one.