POLI 30 D: Political Inquiry

Professor Umberto Mignozzetti
(Based on DSS Materials)

Lecture 10 | Causal Inference with Observational Data I

Before we start

Announcements:

- Quizzes and Participation: On Canvas.
- GitHub page: https://github.com/umbertomig/POLI30Dpublic
- Piazza forum: Not sure what the link is. Ask your TA!
- Note to self: Turn on the mic!

Before we start

Recap: We learned:

- The definitions of theory, scientific theory, and hypotheses.
- Data, datasets, variables, and how to compute means.
- Causal effect, treatments, outcomes, and randomization.
- Sampling, descriptive statistics, and descriptive plots for one variable.
- Correlation between two continuous variables.
- Prediction of a binary and a non-binary variable.

Great job!

- Do you have any questions about these contents?

Plan for Today

- Review: Causation and Randomized Experiments
- Observational Studies
- Confounding Variables or Confounders
  - Why Are Confounders a Problem?
  - Why Don't We Worry About Confounders in Randomized Experiments?
- How Can We Estimate Causal Effects with Observational Data?
  - Interpretation of $\widehat{\beta}$ When X Is the Treatment Variable and Y Is the Outcome Variable

Review: Causation

- To measure causal effects, we need to compare the factual outcome with the counterfactual outcome
  - Fundamental problem: We can never observe counterfactual outcomes
- To estimate causal effects, we must find or create a situation in which the treatment and control groups are **comparable**.
- Only when that assumption is satisfied can we use the *factual* outcome of one group as a good *proxy* for the *counterfactual* outcome of the other.

Review: Randomized Experiments

- In randomized experiments, we can rely on the **random assignment of treatment** to make treatment and control groups, on average, identical
- Thus, we can estimate the average treatment effect with the **difference-in-means estimator**

$$\overline{Y}_{\text{treatment group}} - \overline{Y}_{\text{control group}}$$

Observational Data

- But what happens when we cannot conduct a randomized experiment and have to analyze observational data?
  - *Observational data*: data collected about naturally occurring events (i.e., researchers do not get to assign the treatment)
- We can no longer assume that treatment and control groups are comparable.
- We have to identify any relevant differences between treatment and control groups (known as confounding variables or confounders)
- Then, we have to statistically control for them so that we may claim that the two groups are comparable.

Confounders or Confounding Variables

- A **confounding variable**, or **confounder**, is a variable that affects both:
  1. The likelihood of receiving the treatment X, and
  2. The outcome Y

- In mathematical notation, we represent a potential confounding variable as Z

$$
\begin{tikzcd}
 & Z \arrow[dl] \arrow[dr] & \\
X \arrow[rr] & & Y
\end{tikzcd}
$$

Confounders or Confounding Variables

- Suppose we are interested in the average causal effect of attending a private school instead of a public one on SAT performance.

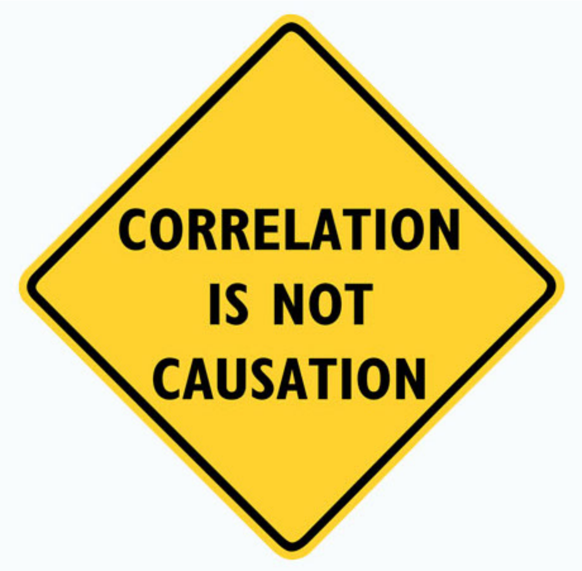
- What is the treatment variable X?
  - What is the outcome variable Y?
  - Can you think of a potential confounder Z?

$$\begin{tikzcd}
& \textit{family wealth} \arrow[dl] \arrow[dr] & \\
\textit{private school} \arrow[rr] & & \textit{test scores}
\end{tikzcd}$$

Why Are Confounders a Problem?

- They obscure the causal relationship between X and Y!
- In the example above, if we observed that, on average, private school students perform better than public school students, we would not know whether it is:
  - Because they attended a private school or
  - Because they came from wealthier families
- We would not know what portion of the observed differences in SAT performance (the difference-in-means estimator), if any, could be attributed to:
  - Attending a private school versus
  - Coming from a wealthy family.

Why Are Confounders a Problem?



Why Are Confounders a Problem?

- In the presence of confounders, correlation does not necessarily imply causation.
- Just because two variables are highly correlated, it does **not** mean that one causes the other:
  - There could be a third variable that causes both!
- Ice cream sales and shark attacks are highly correlated.
  - Does this mean that eating ice cream increases the probability that a shark will attack you?

$$\begin{tikzcd} & \textit{heat} \arrow[dl] \arrow[dr] & \\ \textit{ice cream sales} \arrow[rr] & & \textit{shark attack} \end{tikzcd}$$

Why Are Confounders a Problem?

IN THE PRESENCE OF CONFOUNDERS

- Correlation does not imply causation.

- The treatment and control groups are not **comparable**.

- The difference-in-means estimator does **NOT** provide a valid estimate of the average causal effect!

Why Don't We Worry About Confounders in Randomized Experiments?

- Randomization of treatment assignment eliminates all potential confounders.
    - That is this is the gold standard for causal inference.
- Ensures that treatment and control are comparable by breaking the link between any potential confounder.
- If we have a lottery to randomly determine who will attend the private school, then we break the wealth link.

$$\begin{tikzcd} & & \textit{family wealth} \arrow[dr] & \\ \textit{lottery} \arrow[r] & \textit{private school} \arrow[rr] & & \textit{test scores} \end{tikzcd}$$

How Can We Estimate Causal Effects with Observational Data?

- We cannot rely on random treatment assignments to eliminate potential confounders.
- We must identify all potential confounders and statistically control for them using a multiple linear regression model.
- Before we learn how to do that, we will fit a linear regression model to find the *difference-in-means estimator*.

Using the Linear Regression to Compute the Difference-in-Means Estimator

When X is the treatment variable, and Y is the outcome variable of interest, the estimated slope coefficient ($\widehat{\beta}_1$) is **equivalent** to the *difference-in-means estimator*.

- Let us return to a beloved example: *Does Social Pressure Affect Turnout?*
- Registered voters were randomly assigned to either:
   a. receive a message designed to induce social pressure or
   b. receive nothing

Does Social Pressure Affect Turnout?

(Based on Alan S. Gerber, Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review*, 102 (1): 33-48.)

Does Social Pressure Affect Turnout?

1. Load and look at the data:

```
voting <- read.csv("https://raw.githubusercontent.com/umber
head(voting) # shows first six observations
##   birth message voted
## 1  1981      no     0
## 2  1959      no     1
## 3  1956      no     1
## 4  1939     yes     1
## 5  1968      no     0
## 6  1967      no     0
```

2. Creating the treatment variable:

```
voting$pressure <-  ifelse(voting$message=="yes", 1, 0)
```

Does Social Pressure Affect Turnout?

3. Compute difference-in-means estimator directly

```
mean(voting$voted[voting$pressure==1]) -
  mean(voting$voted[voting$pressure==0])
## [1] 0.08130991
```

4. Alternatively, we can fit a linear model where X is the treatment variable and Y is the outcome variable.

Does Social Pressure Affect Turnout?

- Recall: the R function to fit a linear model is `lm()`

```
lm(voted ~ pressure, data=voting)
##
## Call:
## lm(formula = voted ~ pressure, data = voting)
##
## Coefficients:
## (Intercept)     pressure
##     0.29664      0.08131
```

- Fitted model: $\widehat{voted} = 0.30 + 0.08 \; pressure$
- Note that $\widehat{\beta}_1$ has the same value as the difference-in-means estimator above (both equal 0.08)

Interpretation of $\widehat{\beta}_1$ When X Is the Treatment Variable, and Y Is the Outcome Variable

- Start the same as in predictive models:
  - Definition: $\widehat{\beta}$ is the $\triangle\widehat{Y}$ associated with $\triangle X$=1
    - $\widehat{\beta} = 0.08$ is the $\triangle\widehat{voted}$ associated with $\triangle pressure$=1
  - Receiving a social-pressure inducing message is associated with a predicted increase in the probability of voting of 8 p.p., on average
- Unit of measurement of $\widehat{\beta}_1$? same as $\triangle\overline{Y}$.
  - Since Y is binary, $\triangle\overline{Y}$ is measured in p.p., and so is $\widehat{\beta}$ (after x 100)

Interpretation of $\widehat{\beta}_1$ When X Is the Treatment Variable, and Y Is the Outcome Variable

- Since X is the treatment variable and Y is the outcome variable, $\widehat{\beta}_1$ is equivalent to the difference-in-means estimator
- As a result, we can interpret $\widehat{\beta}_1$ using **causal langauge**
- **Predictive language**: We estimate that receiving the message inducing social pressure ***is associated with a predicted increase*** in the probability of voting of 8 p.p., on average
- **Causal language**: We estimate that receiving the message inducing social pressure ***increases*** the probability of voting by 8 p.p., on average

Interpretation of $\widehat{\beta}_1$ When X Is the Treatment Variable, and Y Is the Outcome Variable

- This should be a valid estimate of the average treatment effect if there are no confounding variables present:
  - If registered voters who received the message are comparable to those who did not.
- Since the data come from a randomized experiment, there should be no confounding variables (why?)
- And thus, the difference-in-means estimator should produce a valid estimate of the average treatment effect

Interpretation of $\widehat{\beta}_1$ When X Is the Treatment Variable, and Y Is the Outcome Variable

- **Conclusion:** A message inducing social pressure increases the probability of voting by eight p.p., on average.
  - Valid estimate of the ATE if registered voters who received the message are comparable to those who did not.
  - This is a reasonable assumption, given that the data come from a randomized experiment.
- Note that this is the same conclusion we arrived at in a previous lecture.

Interpretation of $\widehat{\beta}_1$ When X Is the Treatment Variable, and Y Is the Outcome Variable

INTERPRETATION OF THE ESTIMATED SLOPE COEFFICIENT IN THE SIMPLE LINEAR MODEL:

- By default, we interpret $\widehat{\beta}_1$ using predictive language: It is the $\triangle\widehat{Y}$ *associated with* $\triangle X$=1.
- When X is the treatment variable, then $\widehat{\beta}_1$ is equivalent to the difference-in-means estimator and, thus, we interpret $\widehat{\beta}_1$ using causal language: It is the $\triangle\widehat{Y}$ *caused by* $\triangle X$=1. This causal interpretation is valid if no confounding variables exist: the treatment and control groups are comparable.

Summary

- **Today's Class:**
  - Observational Studies
  - Confounding Variables or Confounders
    - Why Are Confounders a Problem?
    - Why Don't We Worry About Confounders in Randomized Experiments?
  - How Can We Estimate Causal Effects with Observational Data?
  - Interpretation of $\widehat{\beta}_1$ when X is the Treatment Variable, and Y Is the Outcome Variable.

- **Next class**:
  - More Causality with Observational Data:
    - We will use *Multiple Regression* models to control for confounders.

Questions?

See you in the next class!