BRN Intermediate Methods Session 1

1. Introduction to Causal Identification

Inhwan Ko (Univ. of Washington, Seattle)

July 08th, 2020

Contents

1. Logistics

2. Causal Identification - From Idea to Reality

- Four Approaches to Causality
- A Fundamental Problem to Causal Identification (Rubin Causal Model)
- A Secondary Problem to Causal Identification (Random Sampling)

Logistics

- 7pm, Wednesdays, Until 8/26
- Pre-reading materials available for all
- Post-session materials available only for those who attended each week (club good, eh?)
- 1hr 30mins of lecture, 1hr of lab session (R studio!)
- All resources will be distributed via email and Zoom chat attachments

Session Materials

- Angrist & Pischke (2008). Mostly Harmless Econometrics: An Empiricist's Companion.
- Kleiber & Zeileis (2008). Applied Econometrics with R. Springer.
- Cowpertwait & Metcalfe (2009). Introductory Time Series with R.
- Imai (2014). QSS. http://qss.princeton.press/
- Wooldrige (2010). Econometric Analysis of Cross-Sectional and Panel Data. MIT Press.

Causal Identification - From Idea to Reality

Brady (2008) says there are four underlying approaches to causality:

1. Neo-Humean Regularity
2. Counterfactual
3. Manipulation
4. Mechanisms

Let's take a look at these with an example of the effect of smoking

Neo-Humean Regularity

If X and Y co-occur, and co-occur "regularly": causal relationship

- Causal direction: temporal precedence
- Causal power: association (p-value, beta, etc.)

ex) People that smoke more have higher chance of suffering pneumonia ex) More smoking is positively associated with chance of having pneumonia

- Problems?

Counterfactual

If X occurs, then Y occurs; if not, then not

- Finding "control" and "treated" cases: all based on counterfactual logic
- "Natural experiments"; as-if randomness of treatment

ex) Two groups that are similar in age, education, health, etc.: One group that smokes has higher chance of suffering pneumonia than the other that doesn't

- Problems?

Manipulation

If a researcher "gives" X, and Y occurs

- Making counterfactual by researcher's intervention
- Laboratory studies; total randomness of treatment

ex) 100 subjects, 50 told to smoke, 50 not to; those told to smoke showed higher chance of suffering pneumonia than those not to (what?)

- Problems?

Mechanisms

X causes Y "through this particular way"

- Focuses on how X causes Y, rather than whether
- Focuses on theoretical explanation rather than empirical association

ex) A certain chemical component in a cigarette makes lung vulnerable in which it increases the likelihood of pneumonia virus to survive longer inside our body

- Problems?

Validity and Four Approaches to Causality

Roberts et al. (2006) explains validity in a following way:

- Validity: the extent to which a measure accurately represents the concept it claims to measure

ex) Pneumonia (The number of virus cells? The pain a patient feels?); Smoking (The number of cigarettes smoked? The relative frequency?)

- Very important for evaluating the validity of finding: "Is your empirical result consistent with what you've claimed to have found?"

Two types of validity: *external* and *internal*

1. External validity: the ability to apply with confidence the findings of the study to other people and other cases
2. Internal validity: the extent to which a study establishes a trustworthy relationship between X (cause) and Y (effect)

Validity and Four Approaches to Causality

Consider this relationship between validity and four approaches to causality:

1. Neo-Humean Regularity (High external validity, low internal validity)
2. Counterfactual
3. Manipulation
4. Mechanisms (Low external validity, high external validity)

Do you think this makes sense?

A Fundamental Problem to Causal Identification

Let's discuss a fundamental problem to modern studies on causal effects which is called "Rubin Causal Model" (1974). (Though I prefer to call it DORMAMMU!)

Revisit our smoking example:

Question: Does smoking make people less healthy?

To answer this question, let's say we have randomly selected 10,000 subjects and analyze their smoking behavior and health data.

Can this research design help us address the question?

A Fundamental Problem to Causal Identification

Let Y_i be the "potential" health status (outcome) of a person i, and let $D_i = \{0, 1\}$ denote whether a person i is a non-smoker ($D_i = 0$) or a smoker ($D_i = 1$).

Thus, for a person i, potential outcome can be formally written as:

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

An ideal is to find the difference between $Y_{1i} - Y_{0i}$. But this is just an ideal. Why?

A Fundamental Problem to Causal Identification

Our data, which is "observed" outcome, can be written in relevance with potential outcome as follows:

$$\begin{aligned} \textit{observed outcome} = Y_i &= \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} \\ &= Y_{0i} + (Y_{1i} - Y_{0i})D_i \end{aligned}$$

That is, if $D_i = 1$, only Y_{1i} is left on the right-hand side, and if $D_i = 0$, only Y_{0i} remains. This reflects the reality that each individual in our data is either a non-smoker or a smoker. There can be no person who is a non-smoker *and at the same time* a smoker.

A Fundamental Problem to Causal Identification

When we compare the difference in "observed" health status (outcome) between smokers and non-smokers, we will calculate:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$$

However, this includes two subtraction terms which are:

$$\begin{aligned} &E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ &\quad = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &\quad = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \text{ (average treatment effect on the treated)} \\ &\quad + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \text{ (selection bias)} \end{aligned}$$

A Fundamental Problem to Causal Identification

$$
\begin{aligned}
&E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\
&\quad = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\
&\quad = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \text{ (average treatment effect on the treated)} \\
&\quad + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \text{ (selection bias)}
\end{aligned}
$$

- The first subtraction term is an average treatment effect on the treated. This term captures the difference between the health of a smoker, $E[Y_{1i}|D_i = 1]$, and what would have happened to them had they not been a smoker, $E[Y_{0i}|D_i = 1]$. This is our ideal finding.

- However, the second term is the difference in average Y_{0i} between smokers and non-smokers. This captures the reality that what make people smoke (i.e. stress, indirect smoking from peers) can negatively affect their health even before they actually start to smoke. This is called **a selection bias** (aka pre-treatment bias).

Random Assignment Solves the Selection Problem

Random assignment of D_i helps us overcome a selection bias because random assignment makes D_i independent of potential outcomes. Recall that we wanted the below equation:

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \end{aligned}$$

This is because given that D_i is randomly assigned, Y_{0i} and D_i are independent. Therefore, $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$.

Random Assignment Solves the Selection Problem

Now consider this:

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \\ &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E[Y_{1i} - Y_{0i}] \end{aligned}$$

In sum, random assignment allows us to estimate $E[Y_{1i} - Y_{0i}]$ through calculating $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$ which is an only feasible option for us given our data. Remember, we are not Dr. Stranger (DORMAMMU!).

But Random Assignment is Rare - How Can We Even Research?

In reality, a case for random assignment of treatment is hard to find. Mostly we will deal with a data set where units are not randomly assigned with treatment.

In the context of climate change research?

- Democracy (X) and Greenhouse Gas Emissions (Y)
- Trade Dependence (X) and Greenhouse Gas Emissions (Y)
- Democracy (X) and Environmental Policy Stringency (Y)
- Local Fiscal Independence (X) and Renewable Energy Production (Y)
- Whatnot?

What is our best available option?

Another (Relatively Weak) Problem to Causal Identification

Not only treatment is rarely assigned, but our data may not be **representative** of the total population- **random sampling issue**.

Unlike random assignment, random sampling issue is more practical than theoretical.

Random assignment and random sampling are different issues- however, if having a treatment makes an individual easier to be sampled (included in the data), they may mean the same.

When talking about "selection bias," be careful of whether you are refering to one from random assignment issue or random sampling issue. Both issues, however, are serious threats to our endeavor of identifying causality.

So, What Do We Do About It?

The best known solution for mitigating a selection bias when (i) treatment is not randomly assigned and (ii) when we cannot use population data is a multivariate regression. Reasons?

1. Random treatment: we can include possible **confounders** in the model
2. Population data: we can estimate population parameters with sample data

Yet, we must know that regression is based on a wide array of assumptions (i.e. ordinary least squares- based on Gauss-Markov assumption).

Next time, we cover the basics of regression with three questions:
1. How can regression help us estimate population parameters with sample?
2. How can control variables help us mitigate a selection bias?
3. What assumptions are needed to make our estimates unbiased and efficient?