BRN Research Methods Workshop 1

1. Estimating Causal Effects from Observational Data

Inhwan Ko (Univ. of Washington, Seattle)

July 15th, 2021

Contents

1. Four Ways to Understand Causality- Review

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.

1.1. 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

1.2. 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

1.3. 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?)

1.4. 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

2. 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). (DORMAMMU!)

Revisit our smoking example: 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?

2.1. Potential outcome

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

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

$$\textit{potential outcome} = \begin{cases} Y_i^t & \text{if } T_i = 1 \\ Y_i^c & \text{if } T_i = 0 \end{cases}$$

An ideal is to find the difference between $Y_i^t - Y_i^c$. But this is just an ideal. Why?

2.2. Observed outcome

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

$$\begin{aligned} \textit{observed outcome} = Y_i &= \begin{cases} Y_i^t & \text{if } T_i = 1 \\ Y_i^c & \text{if } T_i = 0 \end{cases} \\ &= Y_i^c + (Y_i^t - Y_i^c)T_i \end{aligned}$$

That is, if $T_i = 1$, only Y_i^t is left on the right-hand side, and if $T_i = 0$, only Y_i^c 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. In other words, *an occasional smoker is a smoker.*

2.3. Average treatment effect (ATE)

When we compare the difference in "observed" health status (outcome) between smokers and non-smokers, define an average treatment effect (ATE) of smoking $\bar{\delta}$ as:

$$\bar{\delta} = \bar{Y}^t - \bar{Y}^c$$

Where $\bar{Y}^t$ is an average potential outcome of all individuals *if* they are treated, and $\bar{Y}^c$ is an average potential outcome of all individuals *if* they are controlled.

2.4. Sample average treatment effect (SATE)

Again, however, we cannot observe both potential outcomes of any single individual in our data. Therefore, we end up having:

$$\hat{\bar{\delta}} = \bar{Y}^t_{i \in T} - \bar{Y}^c_{i \in C}$$

Where $\hat{\bar{\delta}}$ is a sample average treatment effect (SATE), $\bar{Y}^t_{i \in T}$ is a sample average outcome of the treated, and $\bar{Y}^c_{i \in C}$ is a sample average outcome of the controlled. This is usually what linear regression gives us.

2.5. Can SATE be ATE?

Any estimators of interest are "consistent" if they approach parameter when sample size is large enough (let's talk more about this in the next session). I will identify a condition for SATE to be consistent so that it approaches ATE.

Before that, we should ask: does the *even* distribution of the treated and the controlled in the population matters for estimating the ATE?

For instance, let's say there is population with the size of 100,000, and only one of them is given a treatment. Can we still estimate ATE?

2.5. Can SATE be ATE? (cont.)

Define w as the proportion of the population that would be assigned to the treatment group. Let $\bar{\delta}_{i \in T}$ equal the ATE of the treated and $\bar{\delta}_{i \in C}$ equal the ATE of the controlled. This gives us:

$$\bar{\delta} = w \ \bar{\delta}_{i \in T} + (1 - w) \ \bar{\delta}_{i \in C}$$

Why?

- For instance, revisit a population with 100,000 individuals where all of them are treated ($w = 1.0$). Then, $\bar{\delta}$ should just be $\bar{\delta}_{i \in T}$. And vice versa when $w = 0$.
- When $0 < w < 1$, since $\bar{\delta}_{i \in T} \neq \bar{\delta}_{i \in C}$, we need to assign a different weight by w to each term and hence a different explanatory power for ATE.

2.5. Can SATE be ATE? (cont.)

Formally, even when $0 < w < 1$, $\bar{\delta} = \bar{Y}^t - \bar{Y}^c$ because:

$$\begin{aligned}
\bar{\delta} &= w\ \bar{\delta}_{i\in T} + (1-w)\ \bar{\delta}_{i\in C} \\
&= w\ (\bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in T}) + (1-w)\ (\bar{Y}^t_{i\in C} - \bar{Y}^c_{i\in C}) \\
&= [w\ \bar{Y}^t_{i\in T} + (1-w)\ \bar{Y}^t_{i\in C}] - [w\ \bar{Y}^c_{i\in T} + (1-w)\ \bar{Y}^c_{i\in C}] \\
&= \bar{Y}^t - \bar{Y}^c
\end{aligned}$$

We have established that the proportion of the treated individuals in population does not matter for our definition of $\bar{\delta}$ as ATE, which is $\bar{Y}^t - \bar{Y}^c$. However, $\bar{\delta}$ does not approach SATE, which is $\bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in C}$.

- What else do we need?

2.5. Can SATE be ATE? (cont.)

The answer: it should be that $\bar{Y}^t_{i\in T} = \bar{Y}^t_{i\in C}$ *and* $\bar{Y}^c_{i\in C} = \bar{Y}^c_{i\in T}$ (two sufficient conditions), so that:

$$\begin{aligned}
\bar{\delta} &= [w\ \bar{Y}^t_{i\in T} + (1-w)\ \bar{Y}^t_{i\in C}] - [w\ \bar{Y}^c_{i\in T} + (1-w)\ \bar{Y}^c_{i\in C}] \\
&= [w\ \bar{Y}^t_{i\in T} + (1-w)\ \bar{Y}^t_{i\in T}] - [w\ \bar{Y}^c_{i\in C} + (1-w)\ \bar{Y}^c_{i\in C}] \\
&= \bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in C} \\
&= \hat{\bar{\delta}}
\end{aligned}$$

Conclusion: ATE converges into SATE when $\bar{Y}^t_{i\in T} = \bar{Y}^t_{i\in C}$ *and* $\bar{Y}^c_{i\in C} = \bar{Y}^c_{i\in T}$, regardless of the treatment proportionality of the population. In other words, the average outcome under the treatment and that under the control should not differ between the treatment and control groups.

2.6. Random assignment of the treatment

Most common ways under the experimental design to meet two earlier sufficient conditions is to *randomly assign* the treatment across samples. This allows that the treatment assignment does not affect the *potential* outcomes.

Formally, $\bar{Y}_i^t = \bar{Y}_{i \in T}^t = \bar{Y}_{i \in C}^t$ and $\bar{Y}_i^c = \bar{Y}_{i \in T}^c = \bar{Y}_{i \in C}^c$.

2.6. Random assignment of the treatment (cont.)

On the flip side, if these conditions are not met, our SATE will have two sources of bias:

$$
\begin{aligned}
\bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in C} &= \bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in C} + \bar{Y}^c_{i\in T} - \bar{Y}^c_{i\in T} \\
&= \bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in T} + [\bar{Y}^c_{i\in T} - \bar{Y}^c_{i\in C}] \\
&= \bar{\delta}_{i\in T} + [\text{baseline difference}] \\
&= \bar{\delta} + [\text{baseline difference}] + \bar{\delta}_{i\in T} - \bar{\delta} \\
&= [ATE] + [\text{baseline difference}] + \bar{\delta}_{i\in T} - w\bar{\delta}_{i\in T} - (1-w)\bar{\delta}_{i\in C} \\
&= [ATE] + [\text{baseline difference}] + (1-w)(\bar{\delta}_{i\in T} - \bar{\delta}_{i\in C}) \\
&= [ATE] + [\text{baseline difference}] + (1-w)[\text{heterogenous treatment}]
\end{aligned}
$$

2.6. Random assignment of the treatment (cont.)

$$\bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in C} = [ATE] + [\text{baseline difference}] + (1-w)[\text{heterogenous treatment}]$$

1) Baseline difference occurs because the potential outcome without treatment can be different between the treatment and control groups. For instance, smokers may already have higher level of stress than non-smokers even if they did smoke from the first place. This is also called *pre-treatment bias*.

2) Heterogenous treatment can be produced because the treatment can interact with unobservable factors that are different between the treatment and control groups. For instance, if smokers have already higher level of stress than non-smokers, smoking makes stress level even higher and therefore further worsen their health status than it would worsen the health status of the non-smokers should they smoke. This is also called *post-treatment bias*.

2.7. An alternative to random assignment: Control?

Observation studies often use several control variables to reduce the bias of their SATE of interest. However, this is not a simple task.

Let z be a stress level which we will use as a control variable. If:

1) z both affects the treatment assignment and the observed outcome, then controlling for it mitigates *pre-treatment bias*. In this case, z is called a *confounder*.

2) z affects the observed outcome but also is affected by the treatment assignment, then controlling for it exacerbates *post-treatment bias.* In this case, z is called a *collider*.

Conclusion: Controlling for covariates is not always good!

2.8. Summary

Going back to our original 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: "Does smoking make people less healthy?

The answer is: NO!

- Random sampling matters less for estimating ATE
- Random treatment assignment is crucial; no correlation between treatment assignment and potential outcome is a sufficient condition for SATE to be a consistent estimator of ATE
- Controlling for covariates helps mitigate pre-treatment bias, but may worsen post-treatment bias
- Must have a good reason when controlling for covariates

Coming up next...

- We will apply this framework for the linear regression
- We will introduce two more qualities along with consistency for the SATE to approach ATE, which are unbiasedness and efficiency
- Probability theory and statistical theory will be used to explain unbiasedness and efficiency