Causal Inference

Causality and Selection Bias: Shortcomings of Observational Analyses

Ken Stiller

Part 2

April 2025

Table of Contents

A hypothetical example

Imagine two students who are interested in getting a very high score on their thesis. They are considering the courses they should take and they are undecided between *Causal Inference* or sticking with *Intro to Statistics*.

Y_i : Thesis score is the outcome variable of interest for unit i.

$$D_i = \begin{cases} 1 & \text{if unit } i \text{ received the treatment (taking Causal inference)} \\ 0 & \text{otherwise.} \end{cases}$$

$$Y_{di} = \begin{cases} Y_{1i} & \text{Potential thesis score for student } i \text{ with Causal Inference} \\ Y_{0i} & \text{Potential thesis score for student } i \text{ without Causal Inference} \end{cases}$$

Q: What is the effect of taking Causal Inference on your thesis score?

Defining the Potential Outcomes

Definition: Treatment

D_i : Indicator of treatment status for unit i

$$D_i = \begin{cases} 1 & \text{if unit } i \text{ received the treatment} \\ 0 & \text{otherwise.} \end{cases}$$

Definition: Observed Outcome

Y_i : Observed outcome variable of interest for unit i. (Realized after the treatment has been assigned)

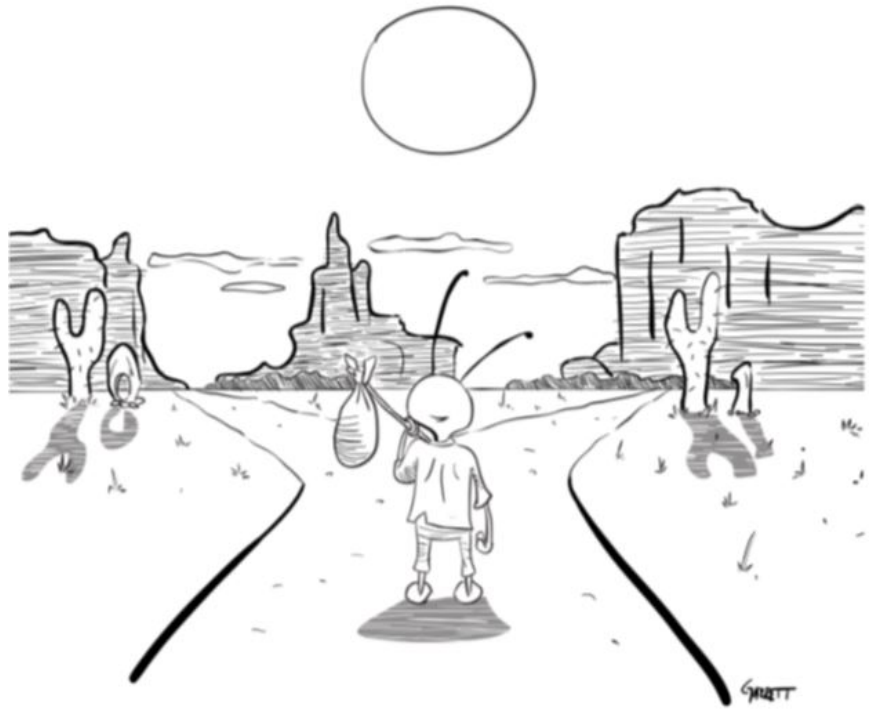
Defining the Potential Outcomes

Definition: Potential Outcomes

Y_{0i} and Y_{1i}: Potential Outcomes for unit i

$$Y_{di} = \begin{cases} Y_{1i} & \text{Potential outcome for unit } i \text{ with treatment} \\ Y_{0i} & \text{Potential outcome for unit } i \text{ without treatment} \end{cases}$$

The Road Not Taken (by Robert Frost)



It shows (1) the actual road that you chose, and (2) the counterfactual road that you could have chosen but did not. **How can we know it made the difference? We don't know what would have happened on the other path.**

Causality with Potential Outcomes

Let D_i denote a binary treatment for unit i, where $D_i \in 0, 1$. Let Y_i represent the observed outcome for unit i. The potential outcomes are thus: Y_{1i}, Y_{0i}

The causal effect of D on Y for i is $\tau_i = Y_{1i} - Y_{0i}$

Definition: Causal Effect

The causal effect of the treatment on the outcome for unit i is the difference between its two potential outcomes:

$$\tau_i = Y_{1i} - Y_{0i}$$

The Fundamental Problem of Causal Inference

The Fundamental Problem of Causal Inference

It is impossible to observe for the same unit i the values $D_i = 1$ and $D_i = 0$ as well as the values Y_{1i} and Y_{0i} and, therefore, it is impossible to observe the effect of D on Y for unit i.

This is why we call this a missing data problem. We cannot observe both potential outcomes, hence we cannot estimate:

$$\tau_i = Y_{1i} - Y_{0i}$$

		Y_{1i}	Y_{0i}
Person 1	Treatment Group ($D = 1$)	Observable as Y	Counterfactual
Person 2	Control Group ($D = 0$)	Counterfactual	Observable as Y

But: We Aim to Make Causal Inference!

In the coming sessions, we will learn ways to address the FPCI.

- Even though it cannot be fully resolved, we can achieve confidence about our findings
- It is absolutely crucial to always be aware what we can claim - as well as about the limitations of our methods

- Now: The limits of linear regressions in the context of the FPCI and why experiments help

Revisiting Linear Regression Models

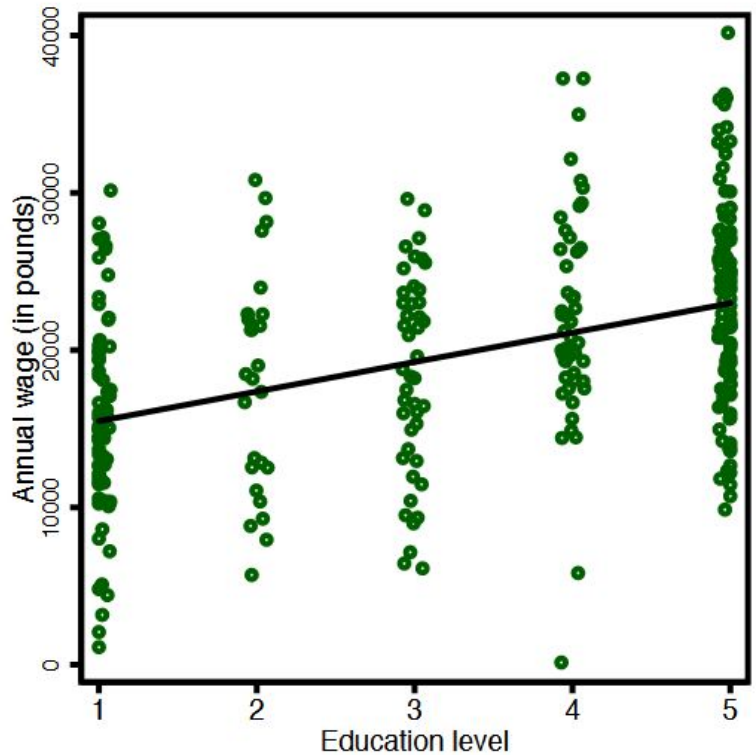
Question: What is the effect of schooling on earnings?

- We obtained some data
- Level of complete education: X
- Annual Income: Y

Linear Equation (OLS Review): The easiest, most parsimonious although not always most adequate, way to summarize the conditional expectation of Y, given X, is to specify a linear model between X and y:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

The Regression Model



- Think of Y as *Salary* and X as *years of schooling*. What we are interested in is β_1. We would like to interpret this coefficient as the *average change produced in individuals' wages by one more year of schooling.*

Linear Regression Model Assumptions (Gauss-Markov Theorem)

$E(Y|X) = \beta_0 + \beta_1 X$

- What does β_0 stand for?
- What does β_1 stand for?
  - Does it matter what the value of X is?

Linear Regression Model Assumptions (Gauss-Markov Theorem)

$$E(Y|X) = \beta_0 + \beta_1 X$$

- What does β_0 stand for?
  - $E(Y|X = 0)$
- What does β_1 stand for?
  - The amount of change in Y in response to a unit change in X
  - Does it matter what the value of X is? No, because there is only one slope.

An Important OLS Assumption

OLS estimation and inference requires various assumptions but we are interested in one of them here. The one that would allow us to interpret β_1 as the average causal effect of education on economic well-being.

The zero-mean assumption

The error term, u_i, is centered around 0 across all values of X: $E[u|X] = 0$. Our X(s) and the error term are *independent*.

Would this assumption hold in our example?

An Important OLS Assumption

This assumption means that all unexplained factors contributing in one's salary have the same value (trivially recoded to zero with the inclusion of a constant) no matter whether one holds a university degree or a high-school degree. This is violated if **omitted variable bias** is present.

- To think whether this assumption is violated we need to think of additional determinants of income. How about parental socioeconomic status (SES)?
- Imagine, then, the true model generating income is the following:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

where $Y = Salary$, $X = Years\ in\ Education$, $Z = ParentalSES$.

The Problem

Since we now know the correct model, we can see what would have happened if we had estimated β_1 from the bivariate regression.

The True Model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

Model without including Parental SES

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Putting the zero-mean assumption into context

The assumption would require the expected parental SES be the same among all people of different educational levels, for those with primary education and those with university degree: $E[u|X] = E[u] = 0$, equivalent to: $E[Z|X] = E[Z]$. This is unlikely to hold, so a model without parental SES (Z) would not estimate the causal effect.

The Selection-on-Observables Assumption

Including Covariates

- Do we really believe that controlling for parental SES is sufficient, i.e. that we know the true model of economic well-being?
- No, this is why we do not only include Parental SES, but other possible predictors of Y.

The reason for doing this is that we want to make the zero-mean assumption **more plausible**.

The Conditioning-on-Observables Assumption

Conditioning on a vector of covariates, Z, we believe that this equality holds: $E[u|X, Z] = E[u|Z] = 0$.

Why do we need causal inference then? The pitfalls of observational research

- Clearly, we don't really know what Z includes. Many plausible candidates might not be measured and might even be measurable. How would one measure ability or aspiration?
- Estimating relationships with controls is certainly better than without them
- The assumption that relationships are monocausal and we have the ability to rule out confounders is elusive.
- We can never account for everything. We just can't!
- **If we are interested in causal links, we need a better, clearer, and stronger framework to understand the social world**

The Fundamental Problem of Causal Inference

The Fundamental Problem of Causal Inference

It is impossible to observe for the same unit i the values $D_i = 1$ and $D_i = 0$ as well as the values Y_{1i} and Y_{0i} and, therefore, it is impossible to observe the effect of D on Y for unit i.

This is why we call this a missing data problem. We cannot observe both potential outcomes, hence we cannot estimate:

$$\tau_i = Y_{1i} - Y_{0i}$$

		Y_{1i}	Y_{0i}
Person 1	Treatment Group ($D = 1$)	Observable as Y	Counterfactual
Person 2	Control Group ($D = 0$)	Counterfactual	Observable as Y

Quantities of Interest

Definition ATE

Average Treatment Effect:

$$\tau_{ATE} = E[Y_1 - Y_0]$$

Definition ATT

Average Treatment Effect of the Treated:

$$\tau_{ATT} = E[Y_1 - Y_0|D = 1]$$

Definition ATC

Average Treatment Effect of the Controls:

$$\tau_{ATC} = E[Y_1 - Y_0|D = 0]$$

An Example: ATE

Imagine a population of 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_i
1	3	1	?	?	?
2	1	1	?	?	?
3	0	0	?	?	?
4	1	0	?	?	?

What is the ATE?

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}]$$

An Example: ATE

Imagine a population of 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_i
1	3	1	3	?	?
2	1	1	1	?	?
3	0	0	?	0	?
4	1	0	?	1	?

What is the ATE?

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}]$$

Since we cannot observe two worlds at the same time, we cannot calculate the ATE.

An Example: ATE

Imagine a population of 4 units with the counterfactual values being made up! (as are all other values in this example!)

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_i
1	3	1	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	1	0	1	1	0

What is the ATE?

An Example: ATE

Imagine a population of 4 units with the counterfactual values are made up! (As all other values in this example!)

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_i
1	3	1	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	1	0	1	1	0

What is the ATE?

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}] = 4/4 = 1$$

An Example: ATE (Continued)

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_i
1	3	1	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	1	0	1	1	0
$E[Y_1]$			1.5		
$E[Y_0]$				0.5	
$E[Y_1 - Y_0]$					1

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}] = 1/4 \cdot (3 + 0 + 1 + 0) = 1$$

An Example: Incorrect ATE

In reality you only get the following:

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_i
1	3	1	3	?	?
2	1	1	1	?	?
3	0	0	?	0	?
4	1	0	?	1	?

Wrong $\tau_{ATE} = E[Y_{1i} - Y_{0i}] = 2 - 0.5 = 1.5$

What is the identification problem?

$$\tau_{ATE} = E[Y_1 - Y_0]$$

$$= \pi(E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1])$$
$$+ \text{(1-}\pi)(E[Y_{1i}|D_i = 0] - E[Y_{0i}|D_i = 0])$$

where π is the share of the treated units in our sample.

What can we observe from the above equation?

1. ?
2. ?
3. ?

What can't we observe from the above equation?

1. ?
2. ?

What is the identification problem?

$$\tau_{ATE} = E[Y_1 - Y_0]$$

$$= \pi(E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1])$$
$$+ \text{(1-}\pi)(E[Y_{1i}|D_i = 0] - E[Y_{0i}|D_i = 0])$$

where π is the share of the treated units in our sample.

What can we observe from the above equation?

1. π
2. $E[Y_{1i}|D_i = 1]$
3. $E[Y_{0i}|D_i = 0]$

What can't we observe from the above equation?

1. $E[Y_{0i}|D_i = 1]$
2. $E[Y_{1i}|D_i = 0]$

Counterfactual outcomes!

What is the identification problem?

The observed difference in the outcome for the treatment and control group are:

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] =$$

$$\mathrm{E}[\mathrm{Y}_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] =$$

$$E[Y_{1i} - Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

What is the identification problem?

The observed difference in the outcome for the treatment and control group are:

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] =$$

$$\mathrm{E}[\mathrm{Y}_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] =$$

$$\underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{ATT} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{SelectionBias}$$

- ATT: Average treatment effect on the treated
- Selection Bias: Differences in the treated and control groups when assigned to the control group.

Both are unobserved and we need to make assumptions!

Key Assumptions for Identification

1. **Conditional Independence Assumption**

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i$$

This suggests that:

$$E[Y_{1i}|D = 1] = E[Y_{0i} + D_i(Y_{1i} - Y_{0i})|D_i = 1] = E[Y_{1i}|D_i = 1] = E[Y_{1i}]$$

This assumptions allows the expectations of the unobservables equal the conditional expectations of the observables for control and treatment. Hence, $E[Y_{1i}|D_i = 1] = E[Y_{1i}]$

Key Assumptions for Identification

2. **Stable unit treatment value assumption (SUTVA)**
   1. Consistency in the treatment group
   2. No Spillover across treatment groups. No interference!

3. **Unconfoundedness**

The causal effect only runs from D_i to Y_i. Potential violations:

   1. Common cause ($D \perp\!\!\!\perp Y$, but Z causes D and Y)
   2. Common effect (selection on the DV; conditioning on a collider; post treatment bias)
   3. Also beware overcontrol bias (D (indirectly) causes Y but it may not appear such if conditioned on an intermediate variable Z)