

# Causal Inference

## Causality and Selection Bias: Shortcomings of Observational Analyses

Ken Stiller

2nd May 2022

# Table of Contents

Recap

Causality & OLS

Potential Outcomes & Identification

Selection Bias

# Definitions

## Causality

Refers to the relationship between events where one set of events (the effects) is a direct consequence of another set of events (the causes). (Hidalgo & Sekhon 2012)

## Causal Inference

The process by which one can use data to make claims about causal relationships. (Hidalgo & Sekhon 2012)

Inferring causal relationships is a central task of science.

## Examples

- ▶ What is the effect of peace-keeping missions on peace?
- ▶ What is the effect of church attendance on social capital?
- ▶ What is the effect of minimum wage on employment?

# A Counterfactual Logic

## Counterfactual Logic

**If X had/had not been the case, Y would/would not have happened**

**Example:** *Does college education increase earnings?*

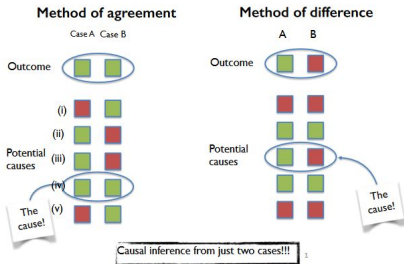
- ▶ If high school grads had instead obtained a college degree, how much would their income change?
- ▶ If college grads had only obtained a high school diploma, how much would their income change?

## John Stuart Mill: A System of Logic (1843)

A series of “canons” for inductive inference. Two well-known methods of comparison:

- ▶ Method of Difference: Units similar in all respects but one manipulable treatment. Units differ in their outcomes.
- ▶ (Method of Indifference: Units dissimilar in all respects but one. Outcome same in both units.)

Suppose all of the potential causes can be enumerated and accurately measured. Then these two methods will under certain conditions tell us the cause of an outcome:



# Mill & the Counterfactual Logic of Causality

## Causal Inference as a counterfactual Problem

Rather than defining causality purely in reference to observable events, counterfactual models define causation in terms of a comparison of observable and unobservable events.

- We need to construct counterfactuals of the observed world as comparison units (Method of Difference).

## Revisiting Linear Regression Models

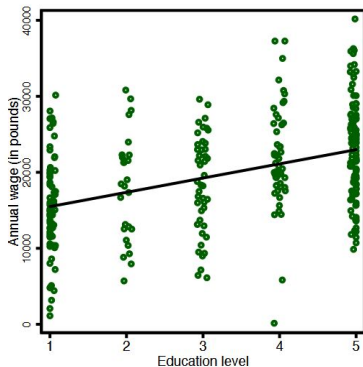
*Question: What is the effect of schooling on earnings?*

- ▶ 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  $\beta_1$ . 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  $\beta_0$  stand for?
- ▶ What does  $\beta_1$  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  $\beta_0$  stand for?
  - ▶  $E(Y|X = 0)$
- ▶ What does  $\beta_1$  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  $\beta_1$  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

**Omitted variable bias:** 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.

- ▶ 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  $Y = \text{Salary}$ ,  $X = \text{Years in Education}$ ,  $Z = \text{ParentalSES}$ .

## The Problem

Since we now know the correct model, we can see what would have happened if we had estimated  $\beta_1$  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 Assumptions

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

## 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**

## 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 choosing *Econometrics I*.

$Y_i$  : 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)

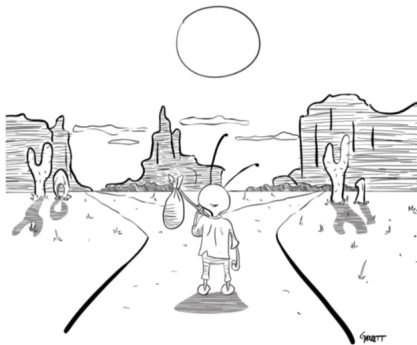
# Defining the Potential Outcomes

## 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

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

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:

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

		$Y_{i1}$	$Y_{i0}$
Person 1	Treatment Group ( $D = 1$ )	Observable as $Y$	<b>Counterfactual</b>
Person 2	Control Group ( $D = 0$ )	<b>Counterfactual</b>	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:

$i$	$Y_i$	$D_i$	$Y_{1i}$	$Y_{0i}$	$\tau_i$
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:

$i$	$Y_i$	$D_i$	$Y_{1i}$	$Y_{0i}$	$\tau_i$
1	3	1	3	?	?
2	1	1	1	?	?
3	0	0	?	0	?
4	1	0	?	1	?

What is the ATE?

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

Because 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 are made up! (As all other values in this example!)

$i$	$Y_i$	$D_i$	$Y_{1i}$	$Y_{0i}$	$\tau_i$
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!)

$i$	$Y_i$	$D_i$	$Y_{1i}$	$Y_{0i}$	$\tau_i$
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)

$i$	$Y_i$	$D_i$	$Y_{1i}$	$Y_{0i}$	$\tau_i$
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:

$i$	$Y_i$	$D_i$	$Y_{1i}$	$Y_{0i}$	$\tau_i$
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?

$$\begin{aligned}\tau_{ATE} &= E[Y_1 - Y_0] \\ &= \pi(E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]) \\ &\quad + (1-\pi)(E[Y_{1i}|D_i = 0] - E[Y_{0i}|D_i = 0])\end{aligned}$$

where  $\pi$  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?

$$\begin{aligned}\tau_{ATE} &= E[Y_1 - Y_0] \\ &= \pi(E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]) \\ &\quad + (1-\pi)(E[Y_{1i}|D_i = 0] - E[Y_{0i}|D_i = 0])\end{aligned}$$

where  $\pi$  is the share of the treated units in our sample.

What can we observe from the above equation?

1.  $\pi$
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:

$$\begin{aligned} & E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] = \\ & E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] = \\ & E[Y_{i1} - Y_{i0}|D_i = 1] + E[Y_{i0}|D_i = 1] - E[Y_{i0}|D_i = 0] \end{aligned}$$

## What is the identification problem?

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

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

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

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

- ▶ 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 assumption allows the expectations of the unobservables equal the conditional expectations of the observables for control and treatment. Hence,  $E[Y_{i1}|D_i=1] = E[Y_{i1}]$

## 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  $Y_i$*

1. Common Cause ( $D \perp\!\!\!\perp Y$ ,  $D$  causes  $Z$ ,  $Z$  causes  $Y$ )
2. Common Effect (selection on the DV, post treatment bias)

## Real world selection bias: health insurance

- ▶ Health insurance in the US used to be voluntary. Obama's "Affordable Care Act" required citizens to buy insurance. Trump rescinded the mandate in 2019. Is mandatory health insurance a good policy? Similarly, is the NHS in the UK a good policy?
- ▶ **Question: What is the effect of health insurance on expenditures and health outcomes?**
- ▶ The US spends a larger fraction of GDP on health care than most other developed countries, but Americans are not necessarily healthier. Health insurance supposedly has health benefits. Yet, health care is expensive and requiring it may be considered to violate freedom.

## What is the Effect of Health Insurance on Health?

We want to compare:

- ▶ The health of someone with insurance.
- ▶ The health of the **the same person** without insurance.

We begin by comparing health outcomes between those with and without insurance.

Should we do this? Are there any potential issues?

# What is the Effect of Health Insurance on Health?

	Some Health Insurance	No Health Insurance	Difference
	A. Health		
Health Index	4.01	3.70	0.31 (0.03)
	B. Characteristics		
Non-White	0.16	0.17	-0.01 (0.01)
Age	43.98	41.26	2.71 (0.29)
Education	14.31	11.56	2.74 (0.10)
Employed	0.92	0.85	0.07 (0.01)
Family Income	106,467	45,656	60,810 (1,355)

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

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] = \underbrace{E[Y_{i1} - Y_{i0}|D_i = 1]}_{ATT} + \underbrace{E[Y_{i0}|D_i = 1] - E[Y_{i0}|D_i = 0]}_{Selection\ Bias}$$

## What is the Effect of Health Insurance on Health?

What can cause a difference in health outcomes for individuals with and without health insurance?

# What is the Effect of Health Insurance on Health?

What can cause a difference in health outcomes for individuals with and without health insurance?

1. Causation: having health insurance directly leads to better health.
2. Reverse causality: the less (or more) healthy are more likely to buy insurance.
3. Confounders: e.g., the more educated tend to buy insurance more often and they know how to live healthier.
4. Any other ideas?

# What is the Effect of Health Insurance on Health?

## Selection bias

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

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

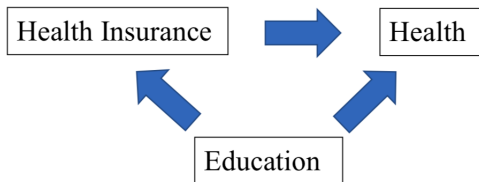
Selection bias is the difference in the potential untreated outcomes between those who were treated and who were not treated. Can we approximate the selection bias here? Is it positive or negative?

$$\underbrace{E[Y_{i0}|D_i = 1] - E[Y_{i0}|D_i = 0]}_{SelectionBias}$$



# What is the Effect of Health Insurance on Health?

## Selection bias



Selection bias is when treatment is assigned in a manner that also affects the outcome. In our example confounders, e.g., education levels, may affect health. Education may also affect the choice to attain insurance. Thus, potential outcomes differ for individuals with and without insurance.