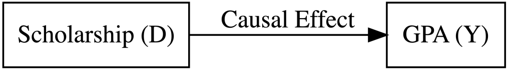
Causal Frameworks I: Exogeneity

Lecture 1 - Introduction to Causal Inference

Kevin Li

What is Causal Inference?

In causal inference, we are interested in how treatment D affects outcome Y.



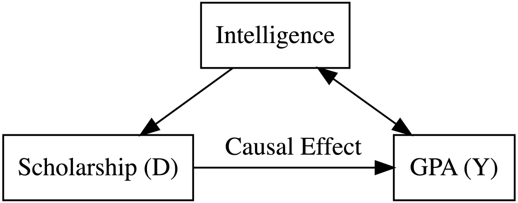
The treatment D is usually assumed to be binary:

$$\mathsf{D_i} = \begin{cases} 1 & \text{if individual i gets scholarship (treated)} \\ 0 & \text{if individual i does not get scholarship (untreated)} \end{cases}$$

How do we find the causal effect? Do we just look at the differences between the GPA of people who got the scholarship vs. people who did not get the scholarship?

Issue: Confounders

What if getting the scholarship is dependent on being smart (such as scoring well on a test). Then, our relationship looks like this:



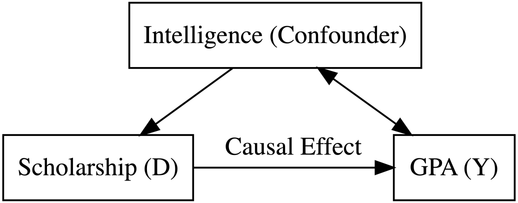
So if we compare the GPA of those with a scholarship and without one, we run into an issue:

Issue: We do not know if differences in the GPA are caused by the scholarship, or caused by differences in intelligence.

Thus, we cannot isolate the effect of the scholarship on GPA.

Formalising Confounders

Intelligence is a confounder.



A confounder X is a variable that meets 3 characteristics:

1. Confounder causes who gets the treatment. In this case, intelligence causes who gets the scholarship.
2. Confounders are correlated with the outcome. In this case, intelligent people tend to get better GPAs.
3. Confounders are not a result of the treatment. In this case, getting a scholarship does not cause intelligence (it's the other way around).

Selection Bias

A confounder X causes/selects different people to get the treatment and not get treatment.

- In our example, intelligent people are more likely to get treatment. So our treated individuals are likely to be more intelligent than our untreated individuals.

Since different people get the treatment than people who do not get the treatment, we do not know if our difference in outcome Y is a result of treatment D, or the pre-existing differences between treated and untreated.

These issue is called **selection bias**.

Solution: Identification

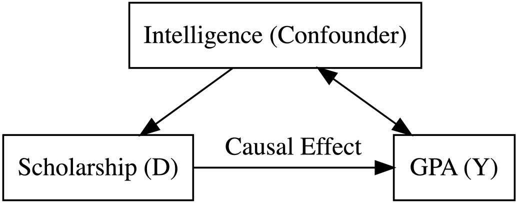
What is the solution to selection bias? Well - we have to account for the influence of confounders in 2 ways:

1. "Control" for confounders through research design - such as finding situations to compare treated vs. untreated where their confounder values are similar.
2. Making **assumptions** for confounders we cannot control for.

This process of getting rid of the influence of confounders is called the **identification** of the causal effect.

What Happens when we Account for Confounders

We know that confounders cause selection into treatment.



- Here - intelligent people are more likely to get treatment (select into treatment), and less intelligent people are less likely to.

Let us assume we somehow **control** for intelligence (how this is done will be the topic of the rest of the classes).

If we control/account for intelligence (and all confounders), now what controls selection into treatment? - **Random Chance**.

Exogeneity

When we control/account for all confounders that cause individuals to select into treatment, the only thing remaining that causes individuals to select into treatment is **randomness**.

This means when all confounders are controlled for, assignment of treatment D is now **random**.

When assignment of any variable (including D) is random, we call this variable **exogeneous**.

- And when there are confounders (and thus not random assignment), we say the variable is **endogenous**.

(You may have heard of exogeneity from Regression - it is the same thing. When we control for all confounders in a regression, our main explanatory variable is now exogenous)

Exogeneity and Causality

We know the following two things:

1. When we control/account for all confounders, we can find the causal effect between treatment D and outcome Y.
2. When we control/account for all confounders, our treatment D is exogenous (randomly assigned).

Thus combining these two statements:

- We can identify/find the causal effect of treatment D on outcome Y, if treatment D is exogenous.

Our goal for causal inference is thus to make our treatment D exogenous (or randomly assigned).

Achieving Exogeneity: Randomisation

We want treatment D to be exogenous/random to identify the causal effects

Well let us personally (as researchers) control who gets treatment through a random number generator (or coin flip). This is called a **randomised controlled trial** or **randomised experiment**.

Downside: we have to be able to control who gets treatment and who does not, so we can assign treatment on some random mechanism.

- It isn't very common where we have control over who gets the treatment, and who doesn't get the treatment.
- Even if it is possible to have control (like a clinical trial), it is also very expensive. So this may not be a feasible strategy.

Achieving Exogeneity: Regression

We can also achieve exogeneity if we include every single possible confounder in our regression with our treatment:

$$Y_i = \alpha + D_i\tau + \mathbf{X}_i^\top\boldsymbol{\beta} + \varepsilon_i$$

- Where $\mathbf{X}_i$ is a vector of values for all confounders for unit i.

If all confounders are in our regression, then τ will be an unbiased estimate of the causal effect of treatment D on outcome Y.

Issue: Often times, we do not know all the confounders, or we do not have data on all the confounders.

- If we miss just one confounder, our causal estimate will be biased. So this may not be a feasible strategy.