The Fundamental Problem of Causal Inference and Common Causes

QTM 110

KEVIN MCALISTER

Administrative Stuff

The first content quiz of the semester is posted on Canvas.

- One week to complete – due on Friday
- No time limit, but don't spend too much time on it
- Shouldn't take more than 30 minutes for a well-prepared student
- Correlation and Discussions of Causation and Counterfactual Dependence

Previously on…

- Causation can be difficult to define.
- We'll adopt the notion of counterfactual dependence/potential outcomes
- In a world where X did not occur, Y would not occur
- A counterfactual is an alternative world where everything was the same **except** for X
- Gives rise to a set of potential outcomes – what would the outcome have been had X been something else?
- The causal effect is then just the difference in the potential outcomes

Counterfactual Dependence

Let X be binary event

- College, drug treatment, batting position

X = 1 “Treated”

- Went to college, Received treatment, Batted left

X=0 “Untreated”

- Didn’t go to college, Didn’t get treatment, Batted right

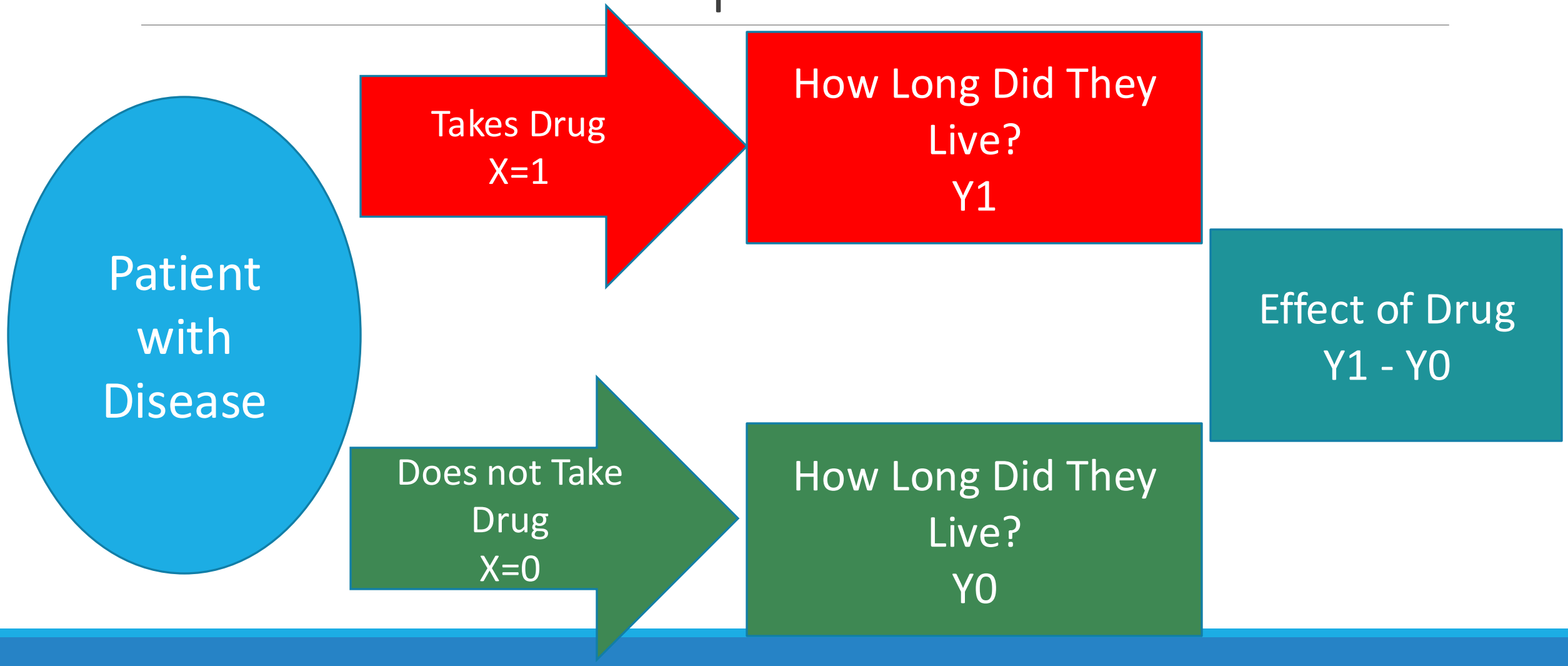
Let Y be some outcome of interest

- Wage, Survival in years, Distance of hit ball

Y1 ≡ Y in the counterfactual world where X = 1

Y0 ≡ Y in the counterfactual world where X = 0

Counterfactual Dependence



Alcoholism and Family Dinners

How many did a DARE program in elementary or middle school?

A nationwide public-service campaign urged parents to eat dinner with their children.

“There is no more important thing a parent can do.”

“The key to ridding our nation of the scourge of substance abuse.”

Eating dinner as a family **causes** a decrease in the chance that a child will use drugs/develop substance dependencies (the claim said this was reduced by 50%!)

Alcoholism and Family Dinners

In groups:

- Clearly outline the causal claim
- What is the causal variable and its possible values?
- What are the potential outcomes?
- Can we realistically observe both the true and counterfactual outcomes? Why or why not? If not, what are some realistic ways we could **approximate** an observable counterfactual world for a child who does not frequently eat dinners with family?
- Use intuition to point out reasons that this claim may be false. In particular, think about the problems you found in the previous part.

Causal Stories

Claim: Smoking causes an increased chance of lung cancer

Story: A person smokes cigarettes for many years. This causes a higher risk of abnormal growths on the persons lungs and, in turn, a higher likelihood of having lung cancer.

Causal story?

Causal Stories

Yes!

Think about a world where a person does not smoke.

No increased chance of developing lung growths and no increase in lung cancer.

Abnormal growths on the lungs are a **mechanism** – a way in which the cause leads to the effect. Without the cause, the mechanism doesn't occur and the effect doesn't happen.

Causal Stories

Claim: Smoking causes an increased chance of lung cancer

Story: A person smokes cigarettes for many years. Studies show that there are certain genetic types that have a predisposition to addiction that also lead to higher chance of developing cellular abnormalities. Therefore, smoking increases the likelihood of having lung cancer.

Causal story? Put another way, can we directly attribute lung cancer to smoking?

Causal Stories

No! (Under the model of counterfactual dependence)

Consider the opposite: In a world where the person doesn't smoke, the genetic predisposition to cellular abnormalities still leads to lung cancer. So, a person still gets lung cancer even if they don't smoke.

It **could even be an extremely strong relationship such that genetics completely determine smoking and cancer**!

Genetics are a **confounder** – genetics is a variable that is correlated with both the causal variable **and** the outcome of interest.

Is this story believable? Does smoking not cause lung cancer?

Causal Stories

We say X causes Y if:

1. X and Y are correlated
2. In a counterfactual world where X did not occur, Y does not occur

Keep it easy:

- Two conditions – Smoke (X = 1) or Don't Smoke (X = 0)
- Outcomes – Lung Cancer (Y = 1) or No Lung Cancer (Y = 0)
- Two **potential outcomes** – Y1 (does the person get lung cancer if they smoke?) and Y0 (does the person get lung cancer if they do not smoke?)
- Effect of smoking on lung cancer – Y1 – Y0

The Fundamental Problem of Causal Inference

How could we figure out which story is true?

The trouble is **observing the counterfactual outcome for an individual**.

- Would she have gotten lung cancer if she didn't have bad genetics? Hold everything else constant.
- Would she have gotten lung cancer if she didn't smoke? Hold everything else constant.

This is what we refer to as the **fundamental problem of causal inference**.

The Fundamental Problem of Causal Inference

The fundamental problem of causal inference underlies **everything** that we will discuss in this class.

It is nearly impossible to see the outcome for the same unit under both possible values of the causal variable.

Examples:

- Smoking?
- Genetics?
- College?
- Mask mandates?

The Fundamental Problem of Causal Inference

Must always be considered when considering individual causal effects.

This is our first occurrence of a fundamental truth:

No causation without comparison/considering the opposite

Without a clear understanding of what is being compared, we cannot say that X causes Y (or that X does not cause Y).

The Fundamental Problem of Causal Inference

Large studies have found that the COVID vaccine is effective for reducing the severity of COVID symptoms.

A causal statement:

"The COVID vaccine causes less severe COVID symptoms and reduces the likelihood that one contracts COVID"

A friend:

"The vaccine doesn't work. I got the vaccine and wore my mask like I was supposed to. I still got COVID and I had a severe fever and trouble breathing."

Does this statement invalidate the vaccine's causal claims?

The Fundamental Problem of Causal Inference

Not really.

Consider the opposite (that we can't actually observe):

What would have happened had he not received the COVID vaccine?

Perhaps he would have ended up in the hospital on a ventilator.
Perhaps he would have had long-lasting lung damage.
Perhaps he would have died.

For my friend's case, we just don't know.

The Fundamental Problem of Causal Inference

Evaluating the efficacy of causal statements without a meaningful counterfactual comparison is a recipe for disaster.

- Drug efficacy
- Government programs
- Surgical approaches
- Voter fraud claims

The list can go on and on and on…

The Fundamental Problem of Causal Inference

What do we do? Is attempting to make causal inferences a fool's errand?

No. We're going to discuss approaches that meet some **assumptions** to get as close as possible to our desired counterfactual world.

The Fundamental Problem of Causal Inference

Approach 1: Be omniscient.

- Not realistic. We can't see everything that is and could be.

Approach 2: Find a single observation that acts as a perfect substitute.

- Me and Identical Twin Me
- Any identical twins in the room?

Approach 3: Find a set of observations that closely approximate a substitute.

- Most likely success is here

The Fundamental Problem of Causal Inference

The big assumption:

There exists some observation (or set of observations) that can actually be observed in the real world that simulates the outcome in the counterfactual world.

Synthetically constructing a substitute for the unobservable counterfactual.

Observable Counterfactuals

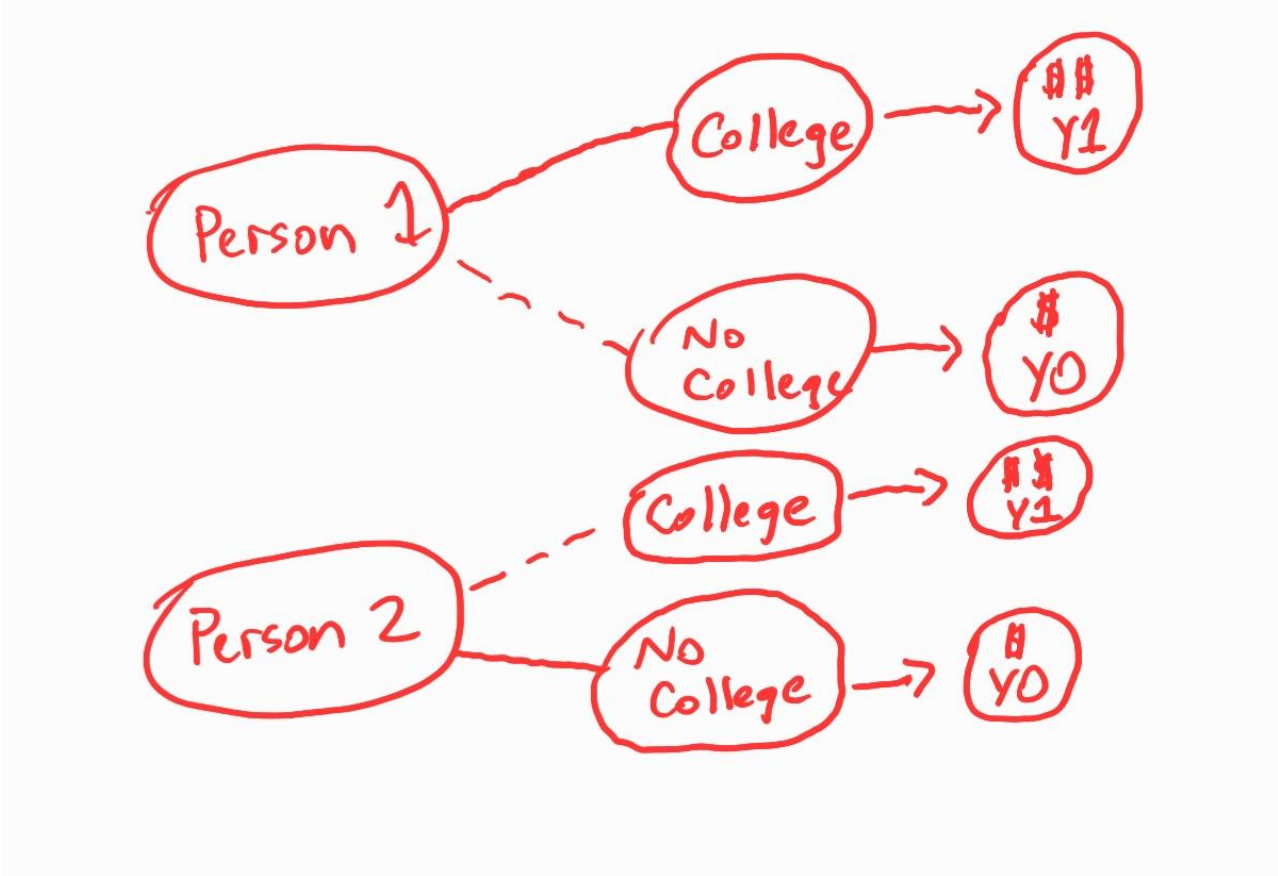
What does it mean to be a perfect substitute?

Causal effect:

$$Y1 - Y0$$

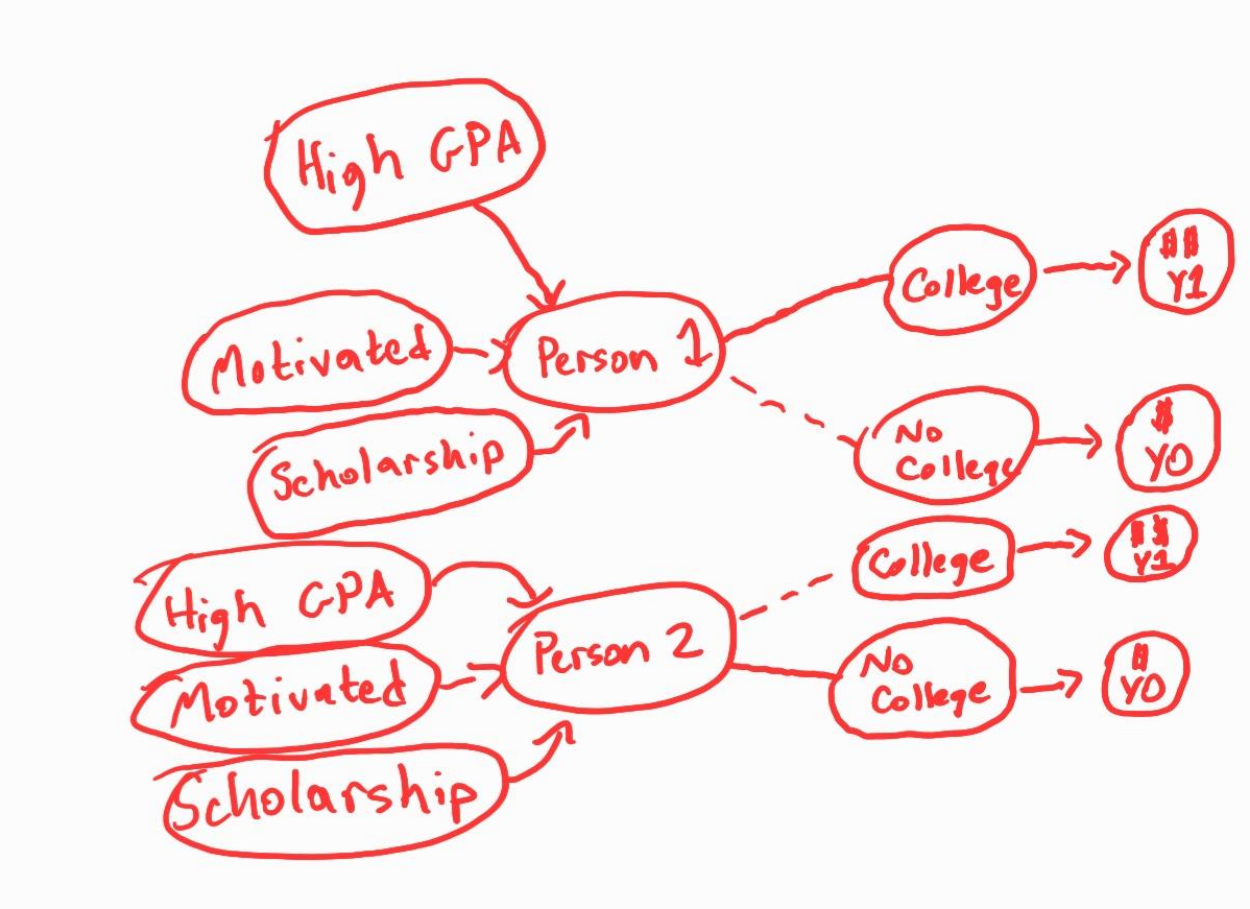
Difference in potential outcomes is a good way to quantify the causal effect.

Observable Counterfactuals



- Consider two different observations
- One goes to college while the other does not
- We observe Y1 when X = 1 and Y0 when X = 0
- If Y1 for person 1 equals Y1 for person 2 **and** Y0 for person 1 equals Y0 for person 2, are they really different for causal effect purposes?
- Equivalence in potential outcomes implies that 1 and 2 are substitutes!
- Problems?

Observable Counterfactuals



- If two people are equivalent in all aspects prior to deciding to go to college, then there's no reason to expect that potential outcomes are different.
- Break an observation into its parts and search for equivalence.
- Gets around the fundamental problem of causal inference by assuming that **equivalence in all pre-treatment factors equals equivalence in potential outcomes**

Observable Counterfactuals

We want to approximate the counterfactual world with an observable observation (or set of observations):

- Equivalence in potential outcomes across groups/individuals gives us counterfactual substitutes
- Fundamental problem of causal inference is still there; can't observe both for each group
- **Assume** that two groups that are the same in all variables other than the causal variable are equivalent in potential outcomes

Coherent and Incoherent Causal Questions

What was the cause of World War II? Why did Kansas City win the Super Bowl?

- Infinite number of factors that, had they been different, would have prevented World War II from being fought.

Proximate Cause?

- Whichever counterfactual world is closest to ours but nonetheless produced no World War II is informative about the most proximate cause of the war.
- Which play was most important?

Solution: Narrow Question

- What the effect of the 1933 election was on European war fatalities over the next two decades.

Answerable and Unanswerable Causal Questions

How could you possibly know what your life would have been like had you not gone to college?

- Many perfectly coherent causal questions may be essentially unanswerable

➔The fundamental problem of causal inference.

- Asking about the effect of the drug for a particular patient is perfectly coherent
  - But we may have no good way of answering that question in a reliable way

Are we on the same page?

Takeaways:

- Causation requires knowing **what would have happened if things were different.**
- A counterfactual world is one where all things are equivalent except for the causal variable of interest (go to college or not)
- We can't **observe** an observation in each counterfactual world – therefore we cannot directly measure all of the potential outcomes
- This is the fundamental problem of causal inference!
- Get around this problem by finding substitutes – equivalent in potential outcomes.

Common Cause

Counterfactuals and Common Cause

Counterfactual substitution requires comparison of like subjects

- Comparison of potential outcomes that differ only in the causal variable

Why does this matter so much?

Counterfactuals and Common Cause

Counterfactual substitution requires comparison of like subjects

- Comparison of potential outcomes that differ only in the causal variable

Why does this matter so much?

- The only way to say that X is causing Y is to see that the only difference between compared groups is in X
- If other important things differ, can we really say that X is the cause?

Counterfactuals and Common Cause

Back to smoking, lung cancer, and genetics

Two groups: those who smoke and those who don't

We observe that those who smoke are 20% more likely to get lung cancer than those who do not.

Correlation.

Counterfactuals and Common Cause

The lurking variable: genetics.

Genetics causes people to have lung cancer, as well. (Correlation)

Counterfactuals and Common Cause

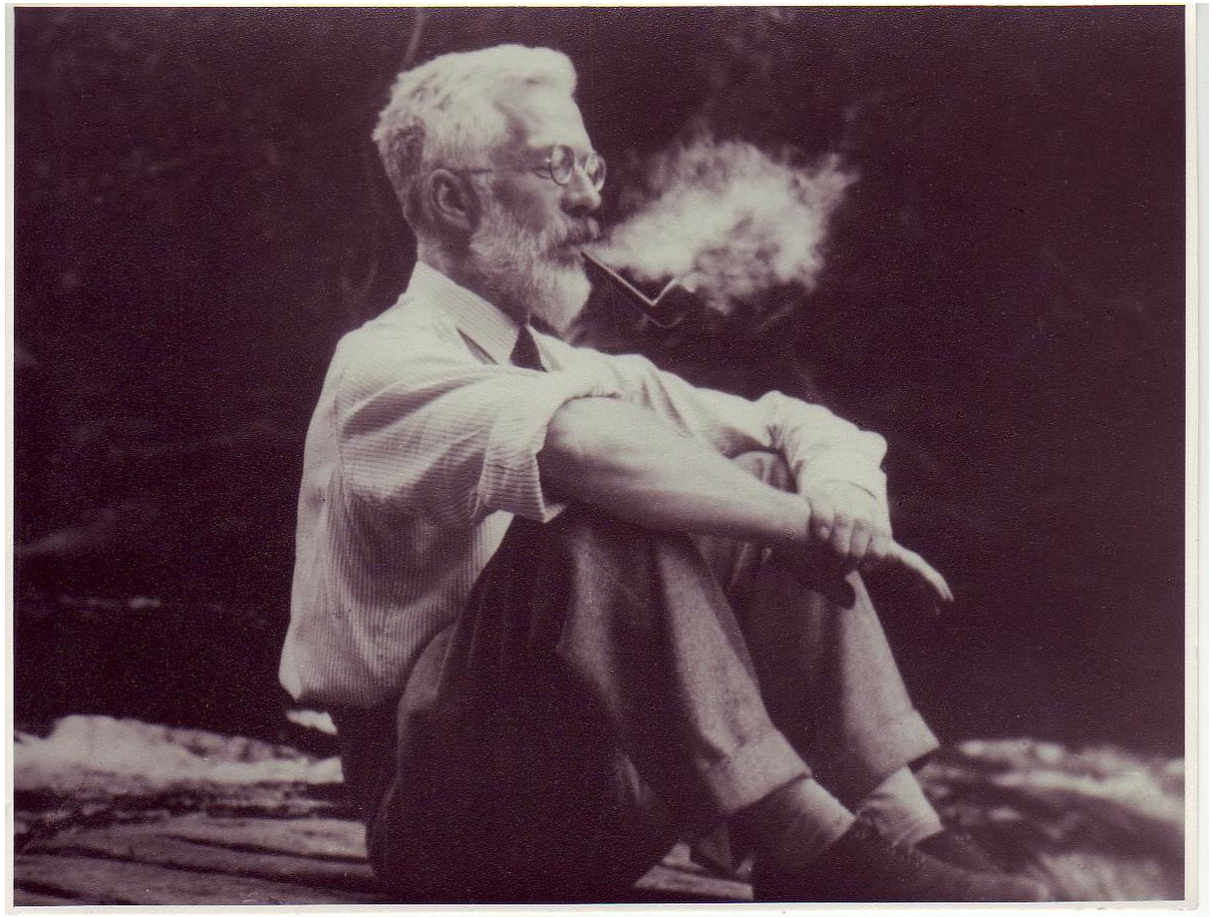
Both can be causes (many things have multiple causes), but how can we **prove** that **smoking** causes lung cancer **separate** from genetics?

Potential outcomes – frequency of lung cancer in the smoking group minus frequency of lung cancer in the non-smoking group

Approach: Proportion of smokers with lung cancer vs. proportion of non-smokers with lung cancer

This difference is positive, so we might be tempted to say that smoking causes lung cancer.

Counterfactuals and Common Cause



- RA Fisher was an important figure in the development of statistics (though his work was the product of a series of problematic agendas…)
- Fisher had a theory about smoking and lung cancer in the face of evidence of a positive correlation
- Some people have a genetic predisposition to be more likely to smoke **and** more likely to get lung cancer
- Fisher theorized that this relationship was so strong that genetics were the main cause of lung cancer – those who smoke are also genetically predisposed to lung cancer. In turn, smoking, itself, wasn't really a cause.
- Fisher also received a large chunk of money from tobacco companies to assist in his research endeavors (I'll let you make of that what you will…)

Counterfactuals and Common Cause

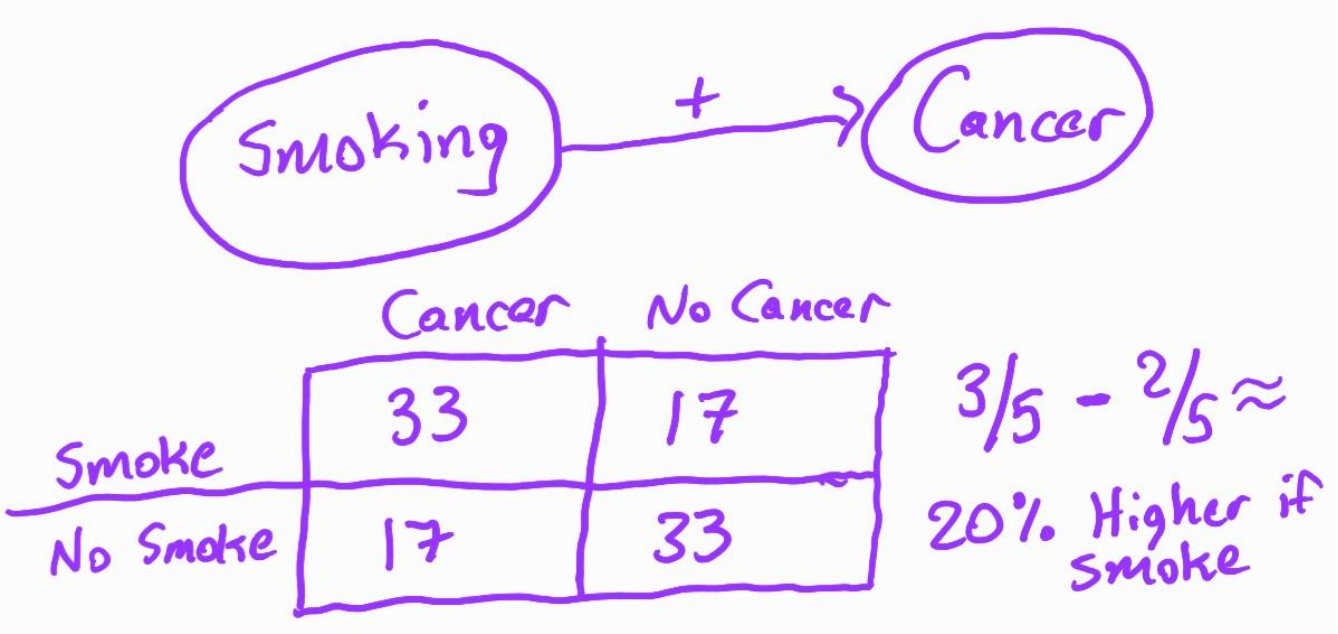
A **confounder** is a variable that correlates with both the causal variable and the outcome.

In the presence of a confounding variable, Z, that is not equivalent across groups, we cannot ever say with absolute confidence that X causes Y because the effect of X on Y is intertwined with the effect of Z on Y.

We could even have the situation where the confounding variable makes the relationship between X and Y look positive when it's actually zero!

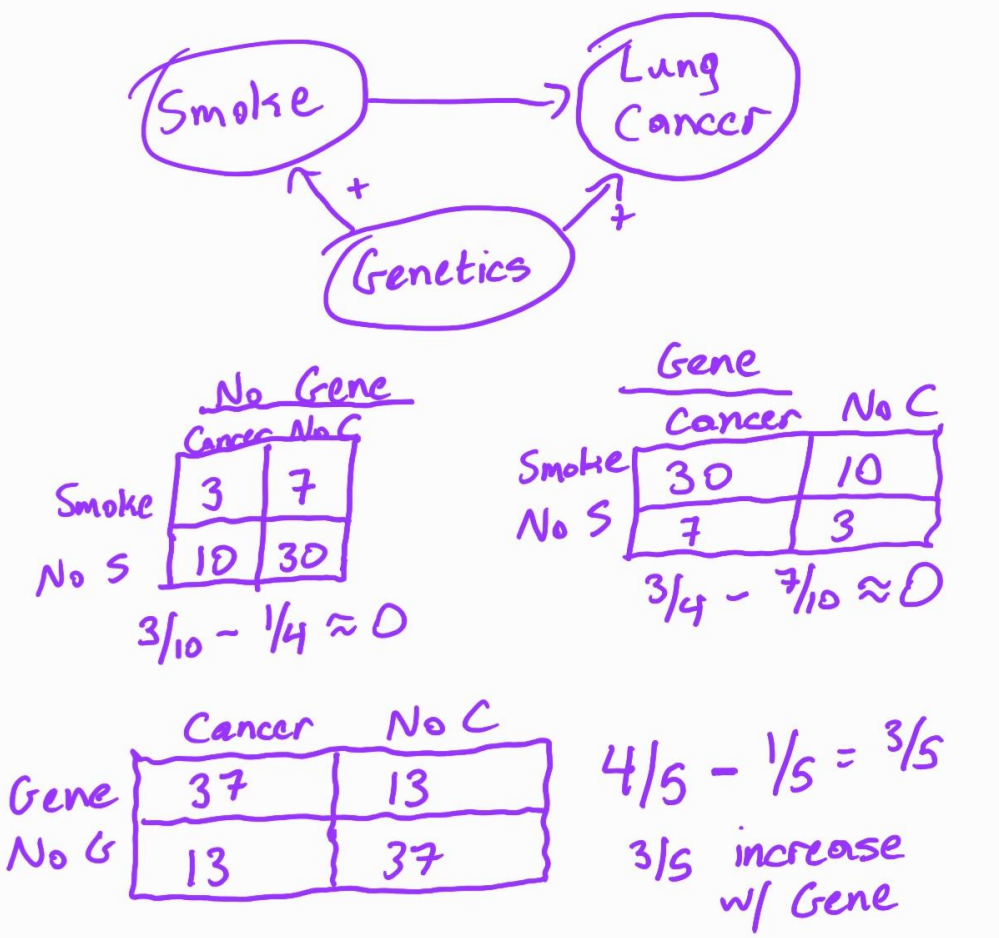
Confounders lead to **bias** - our estimate of the causal effect will never truly be equal to the actual causal effect. More to come...

Counterfactuals and Common Cause



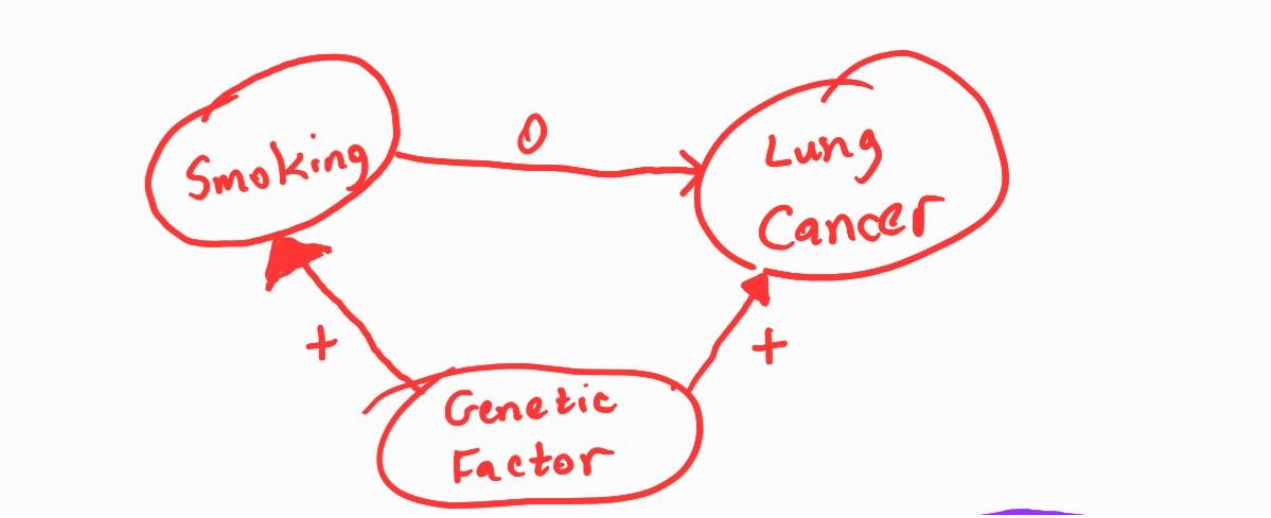
- Let's consider a situation where we collect data about smokers and non-smokers. We see how many get lung cancer.
- Positive correlation between smoking and cancer.

Counterfactuals and Common Cause



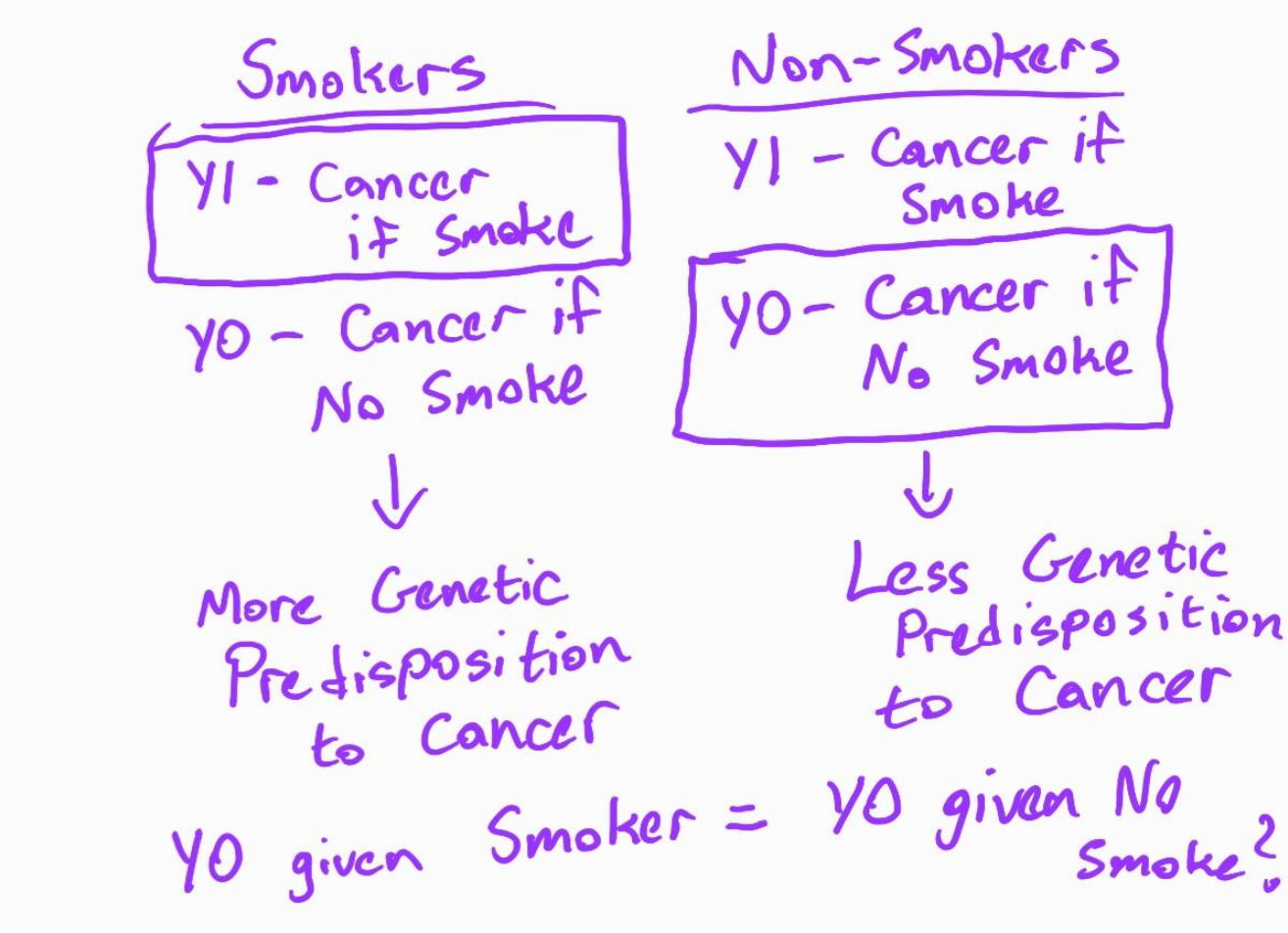
- Data broken out by smoking shows a positive correlation.
- Most people who get cancer have the gene.
- We also see that most people who smoke also have the gene.
- When the effect of smoking on cancer is compared **within** genetic groups, there is a negligible effect.
- But the aggregated data showed that smoking had an effect on cancer!

Counterfactuals and Common Cause



- The smoking group has a higher proportion of individuals with the genetic factor **and** higher rates of lung cancer
- Though smoking, itself, has no effect, we would observe a positive correlation between smoking and cancer.

Counterfactuals and Common Cause



- We can also think of this as a problem of mismatched potential outcomes
- Counterfactual substitute implies that Y1 and Y0 are equal across groups
- In the face of genetics, can we say that Y0 for smokers is the same as Y0 for non-smokers?
- If not, our comparison is not a reasonable counterfactual substitute and we can't plausibly claim counterfactual dependence!

Counterfactuals and Common Cause

Common inferential error:

Common Cause/Confounding

This error arises when a causal claim is made but there is a plausible common cause that is not accounted for in the statement that could be the source of some of the correlation that we see.

Counterfactuals and Common Cause

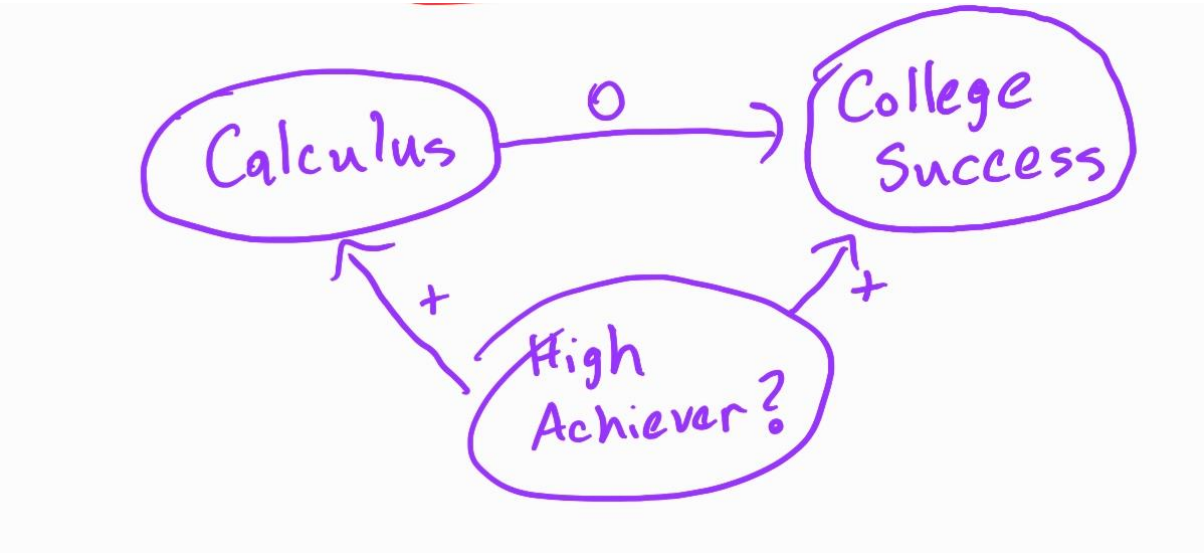
Example: Taking calculus in high school **causes** students to do better in college.

Evidence: How many of you took calculus in high school?

Let's think through this claim.

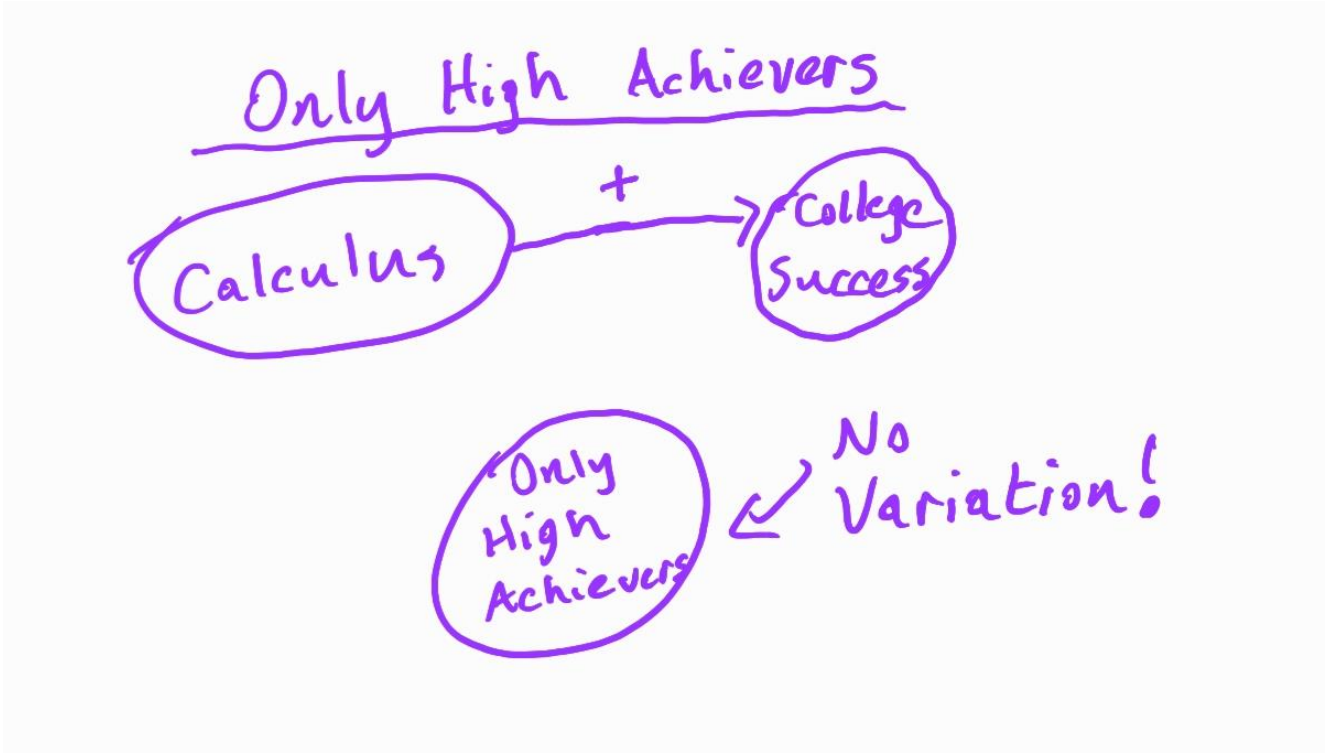
What are some possible common causes that could invalidate this causal claim?

Counterfactuals and Common Cause



- Let's assume that there is a positive correlation between being a high achiever and taking high school calculus **and** a positive correlation between being a high achiever and college success
- As long as we believe that being a high achiever is a confounder, simply looking at data can provide misleading correlations
- Can we rule it out? No.
- Do you think there actually is a causal relationship?
- How could we assess it?

Counterfactuals and Common Cause



- A truth: **correlation requires variation**!
- What if we made sure that the two groups were equivalent in their “achieveness”?
- If we make sure that the two compared groups are only high achievers, there can’t be a correlation between being high achieving and any other variable.
- This is called **controlling** or holding all else constant.
- Each group only differs in whether or not they took calculus in high school!
- More on this to come!

Group Discussion

Claim: “Watching TV as a child causes all sorts of behavioral problems.”

Comparison: Children that Watch TV to Children that Don’t Watch TV

Causal Variable/Treatment: Watching TV

- X = 1 – Treated – Watch TV
- X = 0 – Untreated – Don’t Watch TV
- Outcome: Behavior (measured on some behavioral health index)

Result: Kids who watch TV have worse behavior.

Group Discussion

This is a real claim made by an economics paper in the 1970s!

In your groups:

- Do you buy it? Why or why not?
- Do you think that the evidence supports the causal claim? Think about the counterfactual.
- How could we design a study that would address any problems you found in the previous part?

Takeaways

- Getting around the fundamental problem of causal inference requires finding counterfactual substitutes to observe the unobservable
- Requires comparison of like groups
- When making comparisons, we cannot identify the effect of a single cause if there are **confounders.**
- A confounder is a variable that is correlated with both the causal variable **and** the outcome of interest
- We can deal with confounders by ensuring that any groups that are compared have equivalent levels of a confounding variable (more to come)

Next Time

There are a couple of case studies on small schools posted to the class Canvas site

Please read them before next class – it's a very real situation where policies were created based on bad data literacy!