Unit 6: Causality and Identification

Wooldridge Chapter 3, “Omitted Variable Bias: the Simple Case”, “Omitted Variable Bias: More General Cases” Appendix 3A.4,

Angrist and Pischke Chapter 2.

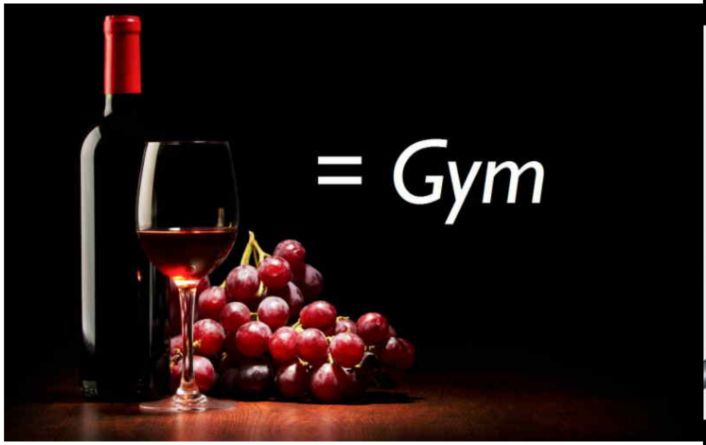
Causality and Identification

Why Study Causality?

- The machinery that we've developed so far applies to any population model, whether that model is causal or not.
  - OLS doesn't care about changes to a specific x – it operates on datasets where the x's are fixed
- But we often have a particular interest in measuring causal relationships.
- We often estimate a model because we want to make a decision or effect a change
  - We want to pull a lever and we want to know what the likely impact will be
    - We might be a government setting a policy
    - Or a company choosing prices
    - Or a doctor deciding whether to prescribe a medication
  - In all these cases, we care about the counterfactual, the state of the world if we made a different decision

This is tough, because we never get to observe the counterfactual, all we can measure are correlations in our population, and as you've heard a million times, correlation is not causation.

Correlation is not Causation



- Is drinking red wine good for your heart?
  - This is a causal question: you might want to know whether to change your drinking habits to stay healthy.
- And there are a number of studies that have shown drinking wine to be associated with lower risk of heart disease.
  - The problem, of course, is that these studies are observational. That is, in the population, people who drink wine have a lower risk of heart disease than people who don't
  - But that doesn't mean that if you drink wine as an individual, you'll lower your risk of heart disease.
  - There are many other reasons that wine drinkers might have a lower risk of heart disease – The rest of their diet might be different, they may spend more time relaxing each day, or spend more time with family. Some of these we can control for in a linear regression, but can we think of all of them?
  - The problem is that drinking wine is an endogenous variable – it's correlated with all these other factors that can also affect risk of heart disease.

Correlation is not Causation

- Is pollution bad for your health?
  - Common sense tells us yes, but how do we know? Can we measure how bad it is?
- If you survey people who live in polluted areas, and people who live in cleaner areas, the people in cleaner areas score higher on health measures.
- But of course the types of people that live there are different.
- Commonly, they are poorer, and poor people have worse health for a wide array of reasons. Diet, exercise, stress, etc…
- So again, the problem is that the pollution you're exposed to is an endogenous variable – it's correlated with many other things that affect health.
- How can you measure how much your health would improve if you actually moved to a less polluted region?

Correlation is not Causation

- To take the point to an extreme: does buying umbrellas cause rain?
- The correlation between the two is quite strong, but of course buying umbrellas is an endogenous variable here, correlated with any atmospheric factors that affect rainfall.

Dealing with Endogeneity

- Our ability to infer causal relationships in the face of endogeneity is quite recent.
- The most important tool we have, instrumental variables, were invented by P.G. Wright in 1928.
  - First used to determine supply and demand curves for linseed oil.
- For a long time, computational limitations were severe, and these methods were usually not feasible.
  - A naïve regression was usually all you needed to publish a paper. (honestly, this is still true in some fields)
- Researchers knew about endogeneity, but it wasn’t regarded as a huge problem.
  - A touch of willful blindness?

The early 80s crisis

- As computing power increased, a lot of fancy techniques were developed through the mid-1900's
  - but they often rested on heroic assumptions.

- The proliferation of naïve regressions and untenable assumptions led to something of a crisis in the early 1980s.
- Edward Leamer (1983):
  - "Hardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analysis seriously."

- Consequences could be severe
  - Isaac Ehrlich on capital punishment in mid 1970s
    - "showed" that capital punishment was a major deterrent using one panel of 35 annual country level observations and one year of national cross section
      - This work became politically influential
    - There was no attempt to deal with endogeneity bias, and the findings have been strongly debated in more recent times.

Randomized Control Trials

- As with many statistical issues, medical science led the way.
- Randomized control trials (RCTs) were developed, in large part, to deal with endogeneity issues.
  - The idea is to make the treatment variable truly random. A truly random variable is uncorrelated with any other random variable, hence exogenous.
    - In essence, we guarrantee that any unmeasured variable is the same, in expectation, for the treatment group and for the control group.
    - So the effects we measure are causal effects
  - The first RCT appeared in a medical journal in 1948, and they became part of official FDA regulations in 1970.

Causal Frameworks for Social Science

- By early 80s there was a sizeable literature bringing ideas from the medical literature surrounding RCTs into the social sciences
  - Donald Rubin, 1970's: Rubin causal model
    - Treatment vs. Control corresponds to two counterfactual states of the world

    "Intuitively, the causal effect of one treatment, E, over another, C, for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to E initiated at t_1 and what would have happened at t_2 if the unit had been exposed to C initiated at t_1: 'If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone,' or because an hour ago I took two aspirins instead of just a glass of water, my headache is now gone.'"

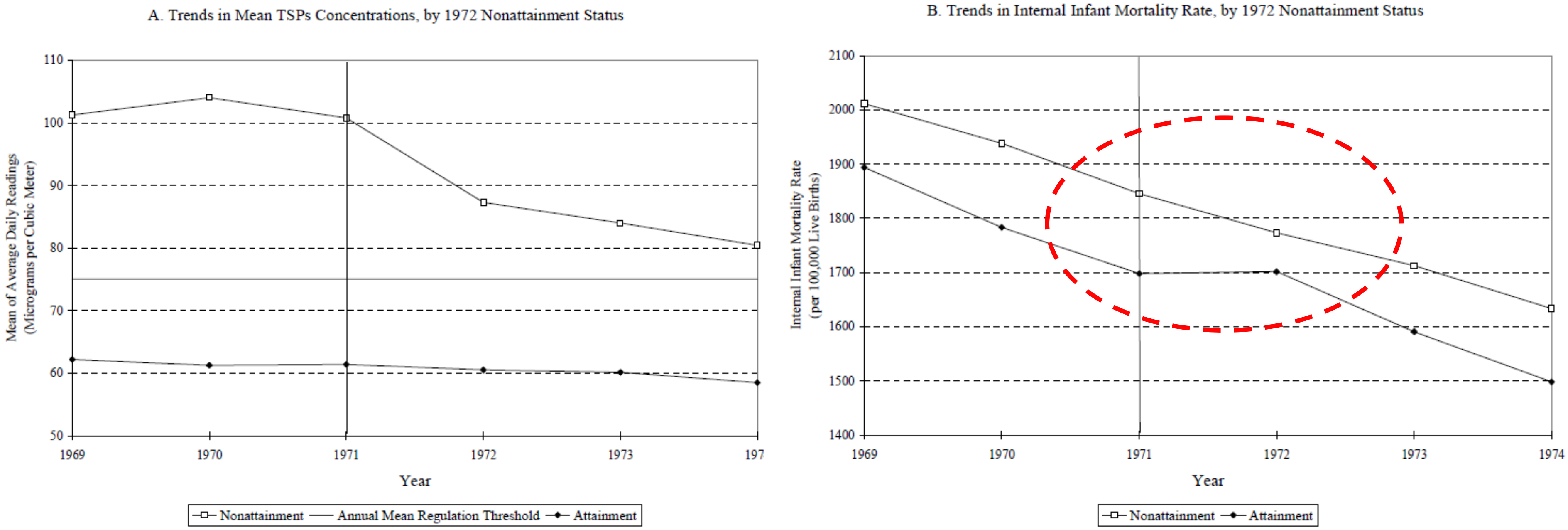
    - So we can talk about causality if we can imagine making different decisions.
    - This interpretation requires that we leave everything else the same other than the decision we make – our physical state, mental state, so forth
  - Rubin's work leads us to a strong emphasis on randomization. We can measure causal effects for decisions that we can treat as random.

The identification revolution

- By the mid 90s, microeconomics in particular had started to realize that if it wanted to be taken seriously it needed to deal with endogeneity
- Specifically, researchers recognized the need for “well-identified” econometrics whenever interpreting something causally
  - “Well- identified” is a term from simultaneous equation modeling.
  - Here, it means that we can correctly “identify” a coefficient in a causal model.

Environmental regulation has both benefits and costs

Chay and Greenstone, 2003
“Air Quality, Infant Mortality, and the Clean Air Act of 1970”



A. Trends in Mean TSPs Concentrations, by 1972 Nonattainment Status

B. Trends in Internal Infant Mortality Rate, by 1972 Nonattainment Status

Source: Authors’ tabulations from EPA’s “Quick Look Reports” data file.

- This study leveraged the clean air act of 1970, which created a set of national standards.
- The change in law affected some states, but not others, depending on previous regulations there.
- The results were causally interpretable because regulation at the county level was functionally random
- “We estimate that a one percent decline in TSPs results in a 0.5 percent decline in the infant mortality rate.”

Donohue and Levitt 2001 –
The Impact of Legalized Abortion on Crime

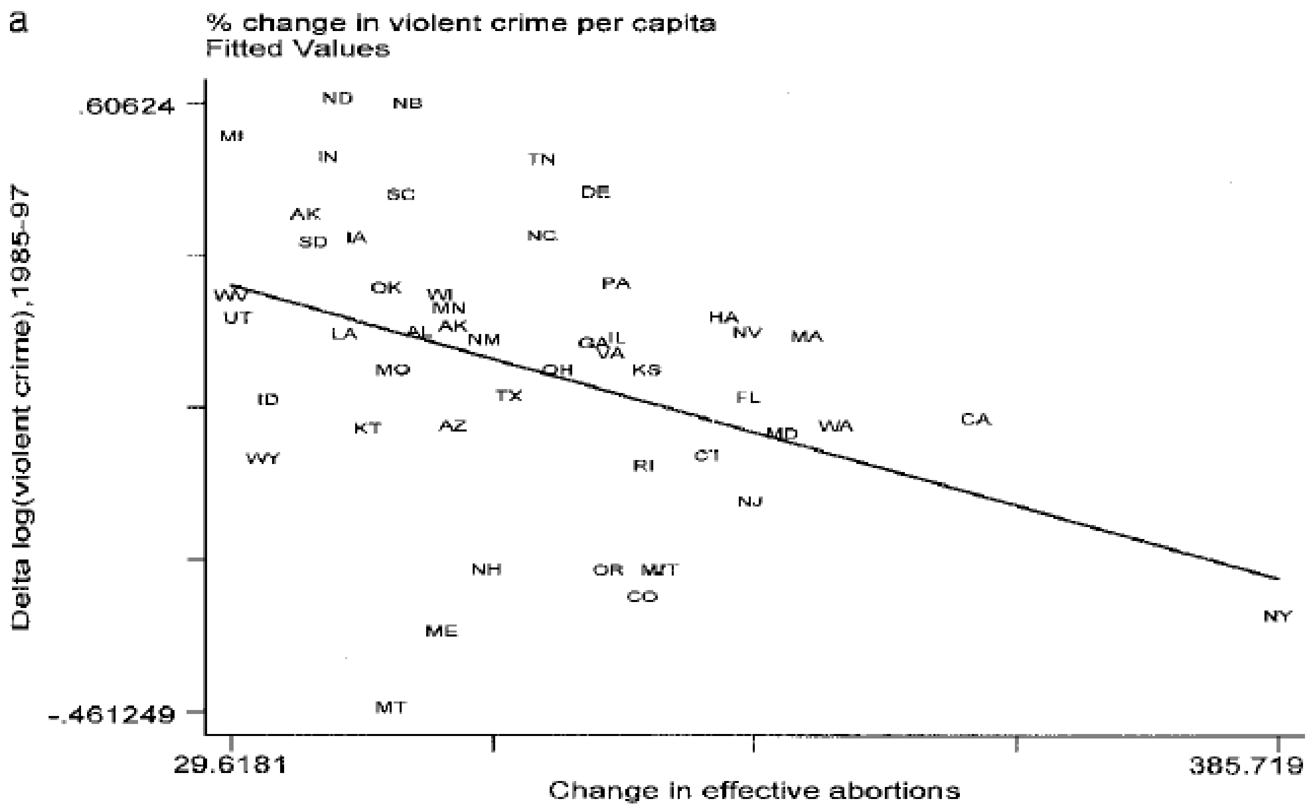


FIGURE IVa

Changes in Violent Crime and Abortion Rates, 1985–1997

- Causally interpretable because regulation of abortion was different at different times in different states and presumably unrelated to murder rates decades later

Looking Forward

- For the rest of this lecture, we’ll delve deeper into causality
- We’ll discuss how we can represent causality, how endogeneity prevents us from identifying the coefficients we want, and ways to control for confounding variables.
- Next week, we’ll cover the workhorse technique social scientists use to overcome endogeneity bias, instrumental variables.

Causal Models

Understanding Causality

- How can we model causal relationships?
- This is a huge question with many competing philosophies.
- Our usual way of thinking about it stems from the counterfactual theories of Neyman, Rubin, and others.
- This is a human-centric approach to causality
  - Some critics would complain that we should be able to reason about causes in the natural world apart from humans
  - But it suffices for most data science purposes in which we want information about decisions.
- We won’t cover the details of causality, but I want to make three main points.

Causal Models

Point 1: Causality is a property of our population model.

- In other words, when we assume our population model is true, we also assume whether it's causal.
  - Now just like any other assumption, this assumption could be realistic or not – we have to build a case for it
- Let's say we our population model is

  $y = \beta_0 + \beta_1 x + u$
- We introduce a manipulation, which is a change in x, dx, and we want to know the corresponding change in y, dy.
- Taking the partial, we have

$$\frac{\partial y}{\partial x} = \beta_1 \quad \text{as long as} \quad \frac{\partial u}{\partial x} = 0$$

- So our coefficient has a causal interpretation as long as the error term doesn't change as we manipulate our x.
- None of these properties are part of the joint distribution of x and y, it's extra structure that we're assuming.
  - The joint distribution is just a static distribution of x's and y's – there's nothing in it about changes to particular datapoint.

Ceteris Paribus

Point 2: Causality is bound up in our ceteris paribus assumption – "all other things equal."

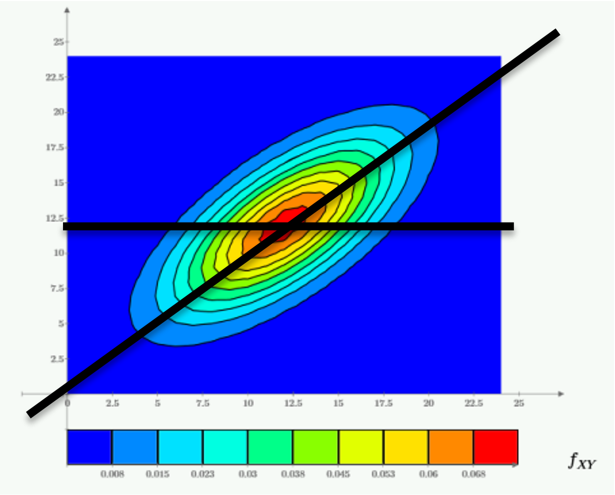
- This means that even though we construct models, we view ourselves as natural scientists
  - We believe that causes are out there in the real world.
  - An outcome that we measure ultimately arises because of some factors.
    - Of course, there may be an infinite number of factors out there, we can only observe a few. Some of them may be random, but at least they exist.
- To make a causal model, we imagine taking every factor except our x's and putting them into the error term of our model.
  - So all the things that we don't actually manipulate, and can't even measure go into the error.
  - With this separation, it's plausible that $\frac{\partial u}{\partial x} = 0$
  - We're just manipulating x, and all other natural factors that affect y are constant.
  - This also means that we'll have to start worrying about what possible factors could be contained in u…

Causality and Exogeneity

- **Point 3:** Causality is not the same as exogeneity.
  - Causality is about whether manipulations to x influence the error term.
  - Exogeneity is about whether OLS can correctly estimate (identity) our coefficients.

Causality and Exogeneity

- Recall this example: Suppose Y represents rainfall and X is umbrella sales.
  - Our joint distribution might look like the heatmap to the right



- We could represent this naively as a population model,

$$y = \beta_0 + \beta_1 x + u$$

With $\beta_1 > 0$ and zero conditional mean, $E(u|x)=0$

  - This model is not causal. There are atmospheric factors that actually affect rainfall, e.g. humidity.
  - we haven't placed these in u. Humidity should be higher to right of the graph.
  - In other words, the model assumes we're manipulating these factors as we manipulate x.
  - On the other hand, zero-conditional mean implies that x is exogenous, and OLS will correctly estimate β_1
- We could also represent this as a causal population model, in which $\beta_1 = 0$ and atmospheric factors are in u. But then x is no longer exogenous.
  - This means that OLS cannot estimate (identify) β_1.

In either case, the coefficients you compute from OLS do not have a causal interpretation.

Omitted Variable Bias in Simple Regression

The lightboard version of this oyster is probably better

Factors in the Error Term

- To write a causal model, we assume that all factors that affect our outcome other than our x variables are contained in our error term, u.
- Part of the reason identification is hard, is that we can't measure all the factors in u.
- We want to estimate our model coefficients, but there's a risk that there's some factor in u that's correlated with x.
  - This is what we call an omitted variable – it's a variable that could affect our outcome but doesn't enter our regression.
  - This could be because we can't measure it.
  - Or it may be a factor we haven't even thought of.
  - When it comes to OLS, what you don't know can hurt you!
- We know that OLS requires x to be uncorrelated with our error term, but what exactly happens if we're wrong?

Omitted Variables

- Let‘s see what happens when we omit a variable from our regression.
- We‘ll start with the simplest case: a regression with one variable and one omitted variable.
- Call the measured variable x_1 and the omitted variable x_2 and suppose the true population model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- Where E(u)=0 and u is uncorrelated with both x‘s.
- If we could measure x_2, OLS would consistently estimate our coefficients.
- But we can‘t measure x_2, so the model we actually estimate is

$$y = \alpha_0 + \alpha_1 x_1 + w$$

- Notice that we changed the names of the coefficients, but could they be the same as the original ones?

Omitted Variables

- Let‘s express x_2 as a function of x_1. We do this by taking a linear regression of x_2 on x_1. (we can always do this)

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

- By the properties of regression, we know cov(x_1,v) = 0.
- If $\delta_1 \neq 0$, x_1 and x_2 are correlated.
- Now plug the expression for x_2 into the true population model and rearrange it to get.

$$\begin{aligned}\Rightarrow \quad y &= \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u \\ &= (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1 + (\beta_2 v + u)\end{aligned}$$

- Since v and u are uncorrelated with x_1 and have expectation 0, the last term is uncorrelated with x_1 and has expectation zero.
- Those are exactly the properties that define the linear regression of y on $x_{1.}$
- Notice that if x_1 and x_2 are uncorrelated, $\delta_1 = 0$ and we get the right coefficient for x_1.
- However, if x_1 and x_2 are correlated, our coefficient is off by $\beta_2\delta_1$
  - this is the omitted variable bias.

Omitted Variable Bias

- Omitted Variable bias is $\beta_2\delta_1$
- It's stronger when
  - the omitted variable has more effect on y.
  - The omitted variable is more correlated with our x_1.
- It's positive when
  - The effect of x_2 on y and cov(x_1,x_2) are both positive or both negative
- It's negative otherwise.
- Sometimes you can use these properties to reason about the direction of bias.

Example

- Let's say I want to convince you that drinking more water is good for weight-loss.
- I collect a sample of people and measure their weight and water intake in cups per day.
- My estimated regression looks like
- Weight = 150 – 1.3 water

  (12) (.02)
- But you realize that drinking water is associated with exercise.
- So imagine that the true population model is actually
- Weight = $\beta_0 + \beta_1$water + $\beta_2$exer + u.
- And exer is related to water according to exer = $\delta_0 + \delta_1$water + v
- Then we can write

  Weight = $(\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)$water + $(\beta_2 v + u)$
- In this case, it's reasonable to assume $\beta_2 < 0$ and $\delta_0 > 0$. This means our omitted variable bias is negative.
- This means it's driving my estimate downward – which is bad since I don't even know if β_1 is negative.
- Sometimes, the opposite happens and you can claim that omitted variable bias is making you estimate more conservative.

Omitted Variable Bias in Multiple Regression

Omitted Variables in Multiple Regression

- Let‘s consider the problem of omitted variables in multiple regression.
- Unfortunately, it‘s harder to derive the direction of bias in this case.
- Say the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$, but we omit x_k
- Let‘s write the regression of x_k on the other independent variables:
- $x_k = \delta_0 + \delta_1 x_1 + \ldots + \delta_{k-1} x_{k-1} + v$.
- And substitute in to get
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k(\delta_0 + \delta_1 x_1 + \ldots + \delta_{k-1} x_{k-1} + v) + u$

  $= (\beta_0 + \beta_k \delta_0) + (\beta_1 + \beta_k \delta_1) x_1 + (\beta_{k-1} + \beta_k \delta_{k-1}) x_{k-1} + (\beta_k v + u)$
- So if we can figure out the signs of β_k and δ_1 we could figure out which way the bias on x_1 goes.
- The sign of β_k is often easy to figure out. But δ_1 can be tricky, because it comes from a multiple regression so it depends on how x_1 is correlated with all the other x‘s.
  - For example, even if x_1 and x_k are positively correlated, δ_1 could be negative.

Example

- Consider the usual wage equation

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

- We know ability is important but we can‘t measure it so it‘s omitted.
- Wooldridge suggests that exper is approximately uncorrelated with educ and abil
- Then the regression of abil on educ and exper becomes simpler. All the variation in educ is unique variation, so the coefficient on educ should be the same as from a simple regression, which we may assume is positive.
- If we further assume that abil is positively correlated with wage, we know that the omitted variable bias is positive on education.
- There are often times when we can make arguments like this to figure out the direction of omitted variable bias – but we have to be careful.

The Experimental Ideal

Experiments

- We've seen than omitted variables create bias in our OLS coefficients.
- Now we're going to start talking about identification strategies.
- An identification strategy is a plan for estimating a causal effect in the face of endogeneity.
- We'll begin with the gold standard for causal claims – true experiments.

- A ***true experiment*** is a study in which the treatment variable is randomly assigned.
- Most often, the treatment is a binary variable. Often, we refer to its levels as treatment and control.
- Random assignment means each person (or other unit) is assigned to treatment or control randomly
  - We can use different rules to assign units to treatment and control
  - The simplest would be to flip a coin for each one.
  - We can increase statistical power if we randomly select half of our sample for treatment and half for control.
  - We can use different rules, but the probability of treatment must not depend on any unmeasured characteristics of each unit of observation.
- A true experiment allows us to estimate causal coefficients consistently, even when we have many unmeasured variables.
- We can see this in a couple of ways.
  - First, a random variable is independent of all other random variables, so it must be uncorrelated with the error term.
  - This immediately tells us that OLS will estimate the treatment effect consistently – that's a basic property.

Randomized Assignment and Omitted Variables

- In more detail, what happens when we have an omitted variable in our error term?
- Here‘s our population model from before where x_2 is omitted.
  - Let‘s assume x_1 is our treatment variable.
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
- $x_2 = \delta_0 + \delta_1 x_1 + v$
- x_2 is an omitted variable so, it appears in u, and the regression we estimate is
- $y = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1 + (\beta_2 v + u)$
- But if x_1 is randomly assigned x_1 and x_2 are independent, so they must be uncorrelated and $\delta_1 = 0$.
- This tells us that omitted variable bias is zero and we correctly estimate β_1
- This is true for every omitted variable that appears in u. Random assignment lets us estimate causal effects consistently.

Potential Outcomes

- Here's another way of looking at random assignment.
- Let's call our treatment variable D_i for individual i.
- For each individual *i*, we imagine that there are two potential outcomes, Y_{0i} and $Y_{1i.}$
  - Y_{0i} is the outcome if the individual doesn't receive the treatment (i.e. D=0)
  - Y_{1i} is the outcome if he or she does receive the treatment (i.e. D=1)
- These potential outcomes are random variables.
- We may think of each of an individuals potential outcomes as being the ***counterfactual*** of the other
- We use these potential outcomes to define a causal effect for subject *i* as a comparison of potential outcomes, $\tau_\iota = Y_{1i} – Y_{0i}$
- Central problem: We can never observe both Y_{0i} and Y_{1i}

Potential Outcomes

- We want to measure E(Y_{1i} – Y_{0i}) the expectation of the treatment effect.
- We can actually find the average population outcomes, E(Y_i |D=1) and E(Y_i|D=0)
- Under random assignment, we can write the expected difference in the population as
- E(Y_{1i} |D=1) – E(Y_{0i}|D=0)
- The key assumption for random assignment is that treatment D is independent of the potential outcomes.
  - This is a neat way of saying that D is independent of all personal characteristics. D can't depend on your outcome if you get the treatment, or your outcome if you don't.
  - It can't be correlated with any personal characteristics that would affect either outcome.
- This assumption lets us drop the conditionals
- E(Y_{1i} |D=1) – E(Y_{0i}|D=0) = E(Y_{1i}) – E(Y_{0i}) = E(Y_{1i} – Y_{0i})
- Even though we only get to observe one outcome for those that get treatment, and one for those that don't, if we subtract the mean of one group from the mean of the other, we get the expected effect for each individual.

What if we can't randomize assignment?

- Randomized experiments are often implausible
  - Cost reasons
    - Even small RCTs are expensive
  - Moral reasons
    - E.g., if you want to study conflict you can't randomly try to provoke violence
  - Logistical reasons
    - We can't randomly assign good institutions, or favorable geography, or social norms, or…
- In such situations we must find some way of inferring (with justification) that our variable of interest can be interpreted as being functionally random
- We basically have two options:
  - Control for confounding variables that are correlated with treatment.
    - We'll talk about this next
  - Find a natural experiment – essentially, this is a scenario in which our treatment variable is being varied by some natural process that's random or plausibly exogenous.
    - This is the topic for next week.

Controlling for Confounders

Controlling for confounders

- If we can’t run a true experiment, we know there could be omitted variables that bias our OLS estimates.
- What can we do about this?

- Most direct approach: measure as many variables as possible and enter them into the regression.
- This is controlling for cofounders.

- Let me say right away that social scientists are generally skeptical of this strategy – if its used by itself
  - Controlling for cofounders is often used in combination with other strategies, and it’s a very important technique.
  - If all you do is control for cofounders, it’s tough to argue that your estimates are unbiased.

- **First problem:** you can’t measure all variables.
  - Famous example: ability
- For variables that we can’t measure, we may try to find a proxy.
  - An example is using IQ as a proxy for ability.
  - A proxy is a variable that is correlated with the omitted variable.
  - Additionally, it should be uncorrelated with the error in the population model
- If both conditions are met, including the proxy will eliminate bias from the omitted variable.
    - Woodridge provides a more detailed explanation.
  - Note that we can’t test the proxy assumptions directly – we have to argue this intuitively.
    - IQ is commonly used, but it’s known to be a rather imperfect proxy for ability.
  - In other fields, proxies are even harder to find.
    - Researchers who study patents would like to measure factors like the novelty of a patent, the breadth of protection, and so forth.
    - But most data comes from the graph of patent citations.
      - For example, the number of citations a patent gets might be used as a proxy for novelty or influence
      - But this is clearly imperfect, and the same graph can only be used for so many factors.
- In short, proxies are a useful tool, but good proxies are hard to find.

Controlling for confounders

- **Second Problem:** there may be confounding factors we haven't even thought of.
  - You might be able to argue that you've accounted for all confounding variables – but those situations are very rare.
    - Usually, it means that treatment is actually random, but the odds of treatment explicitly depend on certain factors.
    - In purely observational studies, it's tough to argue that you've thought of everything that's out there.
    - if you want to know the impact of a marginal year of education, you can include controls for parents' education, wealth, etc., but you can't control for things like an individual's desire to succeed, or the influence of their friends

- **Third problem:** multicollinearity typically increases with more variables
  - As you add variables, you'll be left with less unique variation in your treatment variable.
  - This increases standard errors.
  - For modest datasets (less than a few thousand observations) you typically lose statistical significance as you get to around 10 variables.
  - For country-by-country studies, 5 or 6 variables is often all that you can include before significance disappears.

Controlling for confounders

- **Fourth Problem:** There may be causal pathways from treatment to other variables.
- For example, in the wage model,
- Wage = $\beta_0 + \beta_1$educ + u,
- We might consider including occupation as a control variable
  - Specifically, dummies for various occupations.

– Here's a set of regressions from Angrist and Pischke.
  - Each column is a separate specification – a separate regression
  - As we move from left to right, we're adding control variables.
  - Age dummies don't do too much, the column 3 vars reduce the estimate somewhat
  - In column 4, we've added an aptitude test score. This is a proxy for ability, and we know that ability bias is positive, so no surprise that the estimate decreases a lot.
  - Notice also that standard errors increase as we move right – we're getting more multicollinearity.

	1	2	3	4	5
Controls	None	Age dummies	(2) plus fatheduc, maeduc, race, region	(3) Plus aptitude test score	(4) Plus occupation dummies
Return to educ (std. err.)	.132 (.007)	.131 (.007)	.114 (.007)	.087 (.009)	.066 (.010)

Controlling for confounders

- Finally, in the last column, we've added dummies for different occupations.
  - This might seem like a good idea, since occupation influences wage
  - many people follow the occupations of their parents, or choose their occupations early in life.
- Notice that our estimate has decreased substantially.
  - The problem, is that education influences occupation.
  - If you give someone another year of school, they may choose a different occupation that pays them more money.
  - If we include a dummy for occupation, we eliminate this effect, but we don't want too.
  - In our manipulation, we'd like to alter education and allow it to effect occupation in turn.
- For that reason, we probably shouldn't include occupation in our specification.

	1	2	3	4	5
Controls	None	Age dummies	(2) plus fatheduc, maeduc, race, region	(3) Plus aptitude test score	(4) Plus occupation dummies
Return to educ (std. err.)	.132 (.007)	.131 (.007)	.114 (.007)	.087 (.009)	.066 (.010)

Multiple Specifications

- The occupation example illustrates that there are often tradeoffs when deciding whether to include a variable.
- Should we exclude occupation, in which case we have omitted variable bias?
- Or include it, in which case we increase our errors and lose part of the causal effect we'd like to measure?
- This is part of the reason it's standard practice to include several different specifications when you present regression results.
  - Often, there will be a specification with just the variables of main interest.
  - Then several more with different combinations of covariates.
- This will help reveal whether the effects you measure are robust.
- It will also convince other people that you're not on a fishing expedition, cherry-picking the one specification that shows the effect you're after.

Difference-in-Difference

Difference-in-Difference

- Part of the difficulty with controlling for confounders is that there are potentially so many factors in the error that we can't measure – or that we can't even imagine.
    - It might seem like there's little we can do about things we can't even imagine

- One situation in which we can do a bit better: Suppose we take two measurements of each individual, one at time t=1 and one at time t=2.
- This means we actually have a panel data set.
    - Panel datasets are beyond the scope of this course, but the special case of two time periods is so common it's worth mentioning now.

- Of course, we know that each individual has omitted factors that affect the outcome, both at t=1 and at t=2.
- But we don't look at the outcome.
- Instead, we look at the difference between between y at t=1 and y at t=2, which we write Δy.
- The idea, is that, as long as the effects of the omitted factors are constant, they don't affect the difference.
    - If a person has high ability, for example, their wage will be high in period 1, but also in period 2.
- We can then see how the change in y depends on the treatment.

Diff-in-Diff Example

- Let’s say we measure blood pressure for people on two dates.
- Between the measurements, some people drank red wine, some people didn’t.
  - We know the two groups of people could be very different.
  - There could be many omitted variables correlated with wine drinking.
- Here are the average systolic blood pressure for the two groups of people at both times

	t=1	t=2
Wine drinkers	115	120
Wine non-drinkers	132	128

Diff-in-Diff Example

- The blood pressures look different between the groups, but we don’t know whether that’s because of wine drinking or some omitted variable.
- But let’s look at the change in blood pressure
  - Take the difference from t=1 to t=2
- Now we take the difference between the two values we computed to see what effect drinking wine has
  - hence the name, difference-in-difference.
- In this example, wine drinking increased the change in blood pressure by 9 mg.
  - As long as other factors – exercise, leisure time, etc – that are different between groups only have a constant effect, the diff-in-diff must be caused by the wine itself.

	t=1	**t=2**	**Δy**
Wine drinkers	115	120	5
Wine non-drinkers	132	128	-4
Diff-in-diff			5 – (-4) = 9

Differences-in-Differences

- Let’s see how this works in more detail.
- Here’s a population model for a repeated measures design:
- $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}$
  - y_{it} is the outcome for person i at time t.
  - $d2_t$ is an indicator variable that’s 1 when t = 2.
    - This allows us to have two intercepts, β_0 when t=1 and $\beta_0 + \delta_0$ when t=2
  - a_i is the effect of being individual i. it contains all time-constant factors for the individual.
    - This can be called a fixed effect. It’s fixed through time.
  - u_{it} is the idiosyncratic error for individual i. It there is something about individual i that changes the outcome between t=1 and t=2, it goes in here.

- Write the equation for t=1 and t=2, and subtract one from the other to get
- $\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$

- Now we’re looking at how a change in the treatment affects the change in outcome.
- Notice that a_i has disappeared from the equation.
- To estimate β_1 consistently, we have to assume that Δu_i is uncorrelated with Δx_i
  - Roughly speaking, there are no individual-specific factors that change over time more or less depending on treatment status.

- Diff-in-diff designs are more convincing than merely controlling for covariates.
  - Consider this design whenever you have a treatment that’s sharply defined in time
    - If a policy is enacted in some states but not others
    - If some patients receive a drug but not others
    - And so on.

Recap

Recap

- Today's topic was causality.
- We discussed causal models, and saw that omitted variable bias was an obstacle to estimating causal effects.
- The most convincing way to overcome endogeneity bias is with a true experiment. By randomizing the treatment we remove any correlation with the error term.
- A true experiment isn't always possible, so we looked at other identification strategies – ways to consistently estimate the causal effect of interest.
- We saw that controlling for covariates is an important technique, but one that's generally not convincing by itself.
- A difference-in-difference design removes time-constant individual effects, so helps when we have repeated measures.
- Next week, we turn our attention to natural experiments.
  - A natural experiment is a scenario in which a natural process creates variation that we can exploit to estimate a causal effect.
  - In the simplest case, we find a natural event in which our treatment variable is random, or plausibly exogenous.
    - For example, the vietnam war draft is a classic natural experiment. If we want to study the effects of being drafted, the draft was essentially random (based on birthdays) during that period.
  - More often, the treatment is still endogenous, but we can identify an exogenous source of variation that contributes to the treatment.
    - In this case, we can estimate our effect using instrumental variable methods.
    - These are very important throughout the social sciences and are our topic for next week.