Lecture 13: Measurement Error and Simultaneity

March 27, 2018

Internal Validity

- Before delving into instrumental variables (IV), let's take a step back to other things that can go wrong with internal validity
  - We will find out that IVs solve a lot of these problems
- A taxonomy of what can go wrong:
  1. Omitted Variable Bias: Z such that $\sigma_{X,Z}, \sigma_{U,Z} \neq 0$
     - *Solution:* Control variables (including fixed effects), RCTs, quasi-experimental methods
  2. Specification Bias: The relationship is not linear in X
     - *Solution:* Try logs, polynomials, interactions, etc.
  3. Sample selection Bias: Unrepresentative sample
  4. Measurement error bias
  5. Simultaneity bias: Y and X cause *each other*
- We have discussed (1) at length, solved (2), now let's discuss (3)-(5)

Sample Selection Bias

Definition and Examples

- **Sample Selection Bias** occurs when the sample under consideration is not selected randomly from the population under consideration
- Education and Income example:
    - Only working people have an income
    - So estimated effect of education on income is only the effect on *already employed* persons
    - If education $\Rightarrow$ a higher probability of working, then *total effect* should take into account this probability
- This is a subtle bias (related to external validity in some ways)
    - If sample selection makes $E(U|X) \neq 0$ then OLS is biased
    - If sample selection makes β only consistent for a subpopulation, then interpretation changes

Sample Selection Bias

Solutions (Brief)

- **Sample Selection Bias** occurs when sample under consideration is not selected randomly from population under consideration
- This is *not* selection bias
    - Selection bias was about who is assigned treatment versus control
    - Sample selection is about whether data is *missing* for some group
- **Solutions:**
    - Typically hard to deal with
    - Either get more data from the full population...
    - ... or model the selection issue (hard to do)

Measurement Error (Errors-in-Variables) Bias

Problem: X variable is measured with noise

- Mathematically:

$$\textbf{TRUTH:} Y = \beta_0 + \beta_1 X + U$$
$$\textbf{DATA:} Y = \beta_0 + \beta_1 X^* + U^*$$

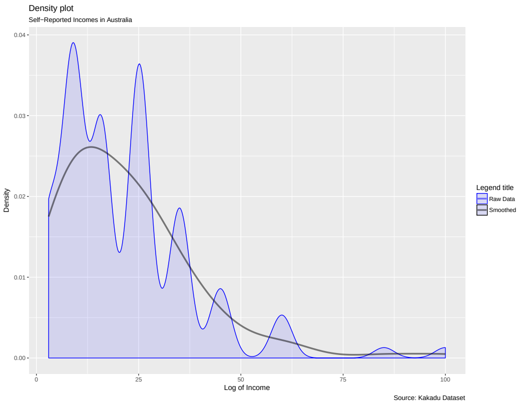
  where X^* is a noisy measure of X

- Examples:
  - Recording errors in data entry
  - Recollection errors in survey data (these are frequent)
  - Rounding errors
  - Hard to measure variables

- Note: we can also have measurement error in Y. Turns out measurement error in Y does not cause any bias (though it does add noise to U, increasing $Var(\hat{\beta}_1)$).

Visualizing Measurement Error: Self-Reported Australian Incomes

Survey question: What is your income in thousands?



- The lumps occur at 5s and 0s
- E.g., people making 41k might say "40" or "45" instead of 41

Measurement Error in X

The Math of Measurement Error

- Classic measurement error: some random noise, v_i, is added to X_i:
  - Formally, v_i is independent of X_i and u_i
  - **Note:** If u_i and X_i are correlated this is very complicated

$$
\begin{aligned}
Y_i =& \beta_0 + \beta_1 X_i^* + u_i \\
Y_i =& \beta_0 + \beta_1 (X_i + v_i) + u_i \\
=& \beta_0 + \beta_1 X_i + \underbrace{\beta_1 v_i + u_i}_{\text{“New” Error Term}}
\end{aligned}
$$

- We will see that this will "attenuate" $\hat{\beta}_1$ toward zero

Classical Measurement Error Math

Suppose that we do not observe X_i, but rather X_i^*, where $X_i^* = X_i + v_i$ and $v_i \sim \mathcal{N}(0, \sigma_v^2)$ and is independent of X_i and u_i. Find the bias when you run the OLS regression:

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

Workspace

Measurement Error in X

- We found that:

$$\hat{\beta}^{OLS} = \beta_1 \times \underbrace{\frac{\sigma_X^2}{\sigma_X^2 + \sigma_v^2}}_{\text{Attenuation}}$$

- Estimated coefficient converges to the truth times an attenuation term
  - Attenuation term is less than 1 $\Rightarrow$ pushes coefficient towards zero
    - It does *not* make the term more negative—it squeezes the coefficient and preserves the sign
  - This "attenuates" the effect of X on Y. Hence the name: **attenuation bias**

Measurement Error in X

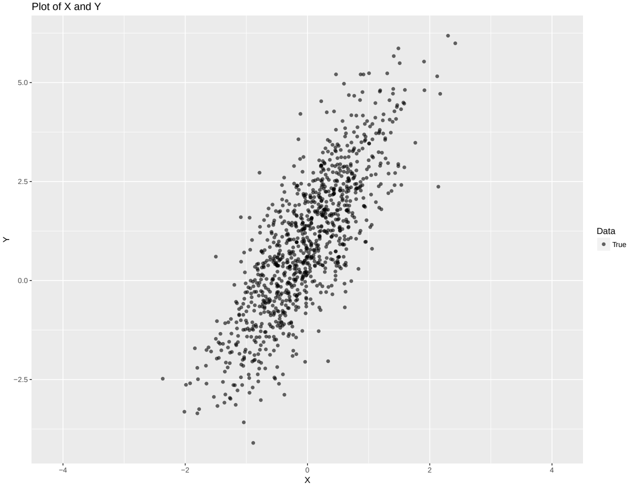
Classical Measurement Error Intuition

- Why does measurement error attenuate the estimated effect of X on Y?
  - Noise in measured X makes it more likely to see high X and low X with some Y because of randomness
  - Extreme example: fix X and keep adding noise—eventually X^* will look like noise itself
- Notice the following rearranging of the attenuation term:

$$\frac{\sigma_X^2}{\sigma_X^2 + \sigma_v^2} = \frac{1}{1 + (\sigma_v^2/\sigma_X^2)}$$

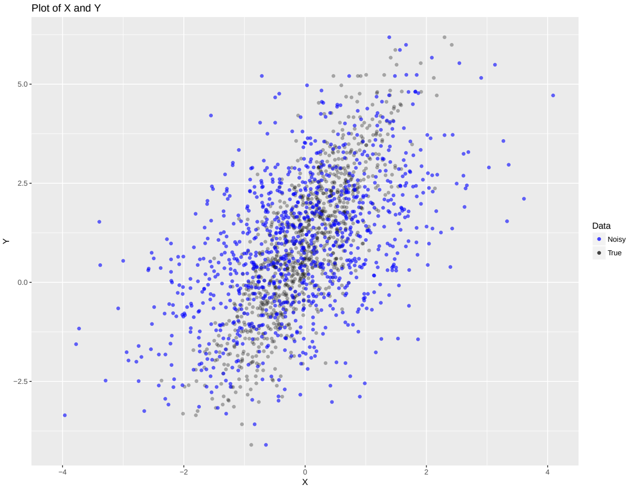
  - σ_X^2/σ_v^2 is called the **signal-to-noise ratio**
  - Larger signal to noise ratio $\Rightarrow$ smaller bias
  - What matter is the *relative* variance of X to v
  - If σ_v^2 is small *relative* to σ_X^2 then attenuation bias isn't too large

Visualizing Measurement Error



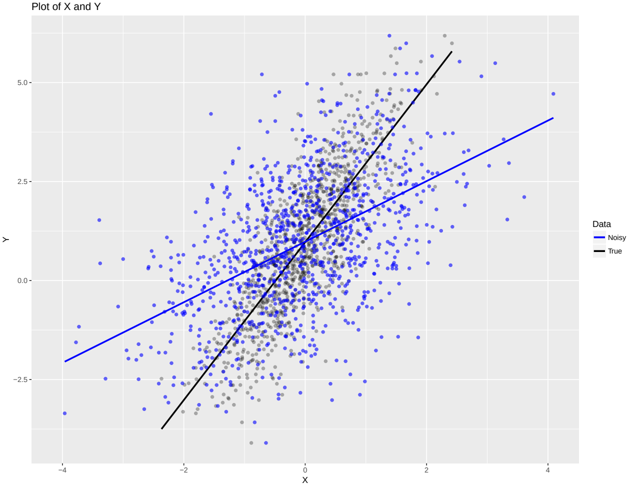
- Suppose this is the true data. But...

Visualizing Measurement Error



- ... you only observe the noisy data...

Visualizing Measurement Error



- ... then the estimated slope is flatter

Simultaneity Bias

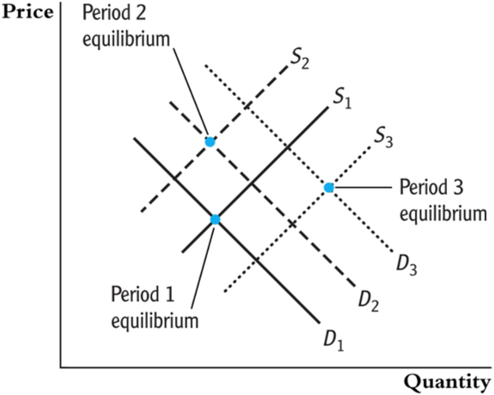
- Suppose we want to estimate the demand elasticity for butter:

$$ln(Q_i^{butter}) = \beta_0 + \beta_1 ln(P_i^{butter}) + u_i$$

- β_1 here is price elasticity of butter
  - e.g., percent change in quantity for a 1% change in price (recall log-log specification)
- The above OLS regression will suffer from simultaneous causality bias

Simultaneity Bias Continued

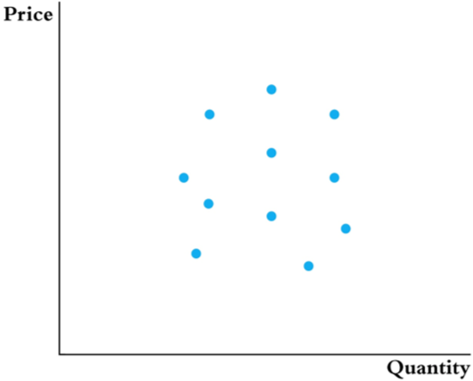
- Simultaneity bas arises because price and quantity are determined *jointly*
- Remember ECO 100:



(a) Demand and supply in three time periods

Visualizing Simultaneity Bias

- So your data will end up looking like this:

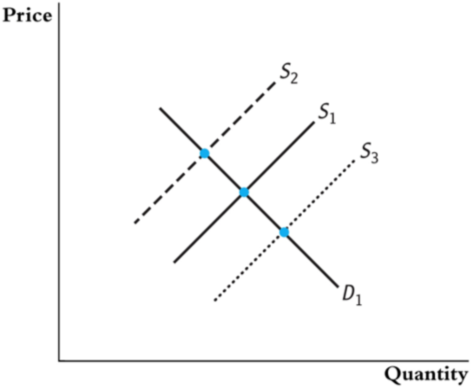


(b) Equilibrium price and quantity for 11 time periods

- You will effectively get a slope of zero in this regression

Visualizing Simultaneity Bias

- But if only supply shifted you would get the correct slope:



(c) Equilibrium price and quantity when only the supply curve shifts

- We thus needs something that only shifts supply (or demand) to solve this problem

Reverse Causality/Simultaneity Bias

- Reverse causality and simultaneity bias are similar ideas
- Simultaneity: X and Y are determined at the same time
- Reverse causality: Y causes X
  - Simple solution: Change your Y and X variable!
- Bigger problem: reverse causality is usually a sign of simultaneity bias
  - Example: More police causes lowers crime, but more crime also causes higher police presence
  - Besides trivial case where you mixed up your X and Y, reverse causality and simultaneity are usually the same thing

The Miracle Worker!

- Instrumental Variables have the ability to solve:
  1. Simultaneity bias (this is what they were originally used for)
  2. Measurement error
  3. Omitted variable bias

Lecture 14: Instrumental Variables

March 27, 2018

Instrumental Variables Introduction

- Three quasi-experimental designs:
  1. Difference-in-Differences
  2. Instrumental Variables
  3. Regression Discontinuity
- Instrumental variables are the most wide-ranging of our quasi-experimental designs, in that they can solve:
  1. Simultaneity bias (this is what they were originally used for)
  2. Measurement error
  3. Omitted variable bias
- We will mostly discuss how we use them to eliminate omitted variable bias

Basic Idea

- Basic Idea of Instrumental Variable (IV):
  - What if we have a variable that is correlated with X *but not with* Y
  - Then any changes in Y caused by that variable will reflect causal changes by X
- Equation of interest:

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

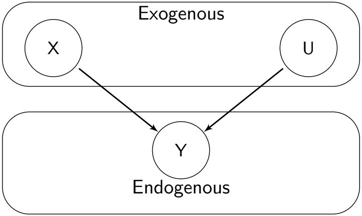
- IVs break X into two pieces that are themselves uncorrelated:

$$X_i = Z_i + V_i$$

  - A piece that is not correlated with U ($Cov(Z, U) = 0$)
  - A piece that is correlated with U ($Cov(V, U) \neq 0$)
  - Finally, $Cov(Z, V) = 0$
- Z_i is an **instrumental variable**

Terminology Review

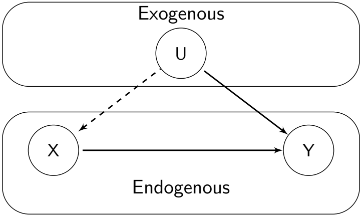
A "Good" Regression:



- **Exogenous Variables:** Variables in the data that do not cause each other
  - U is always exogenous, so exogenous also just means variables not correlated with U
- **Endogenous Variables:** Variables that are determined by exogenous variables in the model
  - U is always in Y so Y is always endogenous

Omitted Variable Bias with Pictures

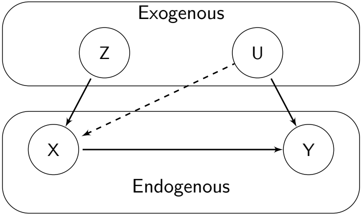
Selection/OVB:



- In this picture X is **endogenous** because U now causes X as well
- What if there is another exogenous variable that *does not* directly cause Y?

Omitted Variable Bias with Pictures

Instrumental Variable:



- Z causes Y *only indirectly* through X
- We can estimate "causal effect" of Z on Y and this MUST be the causal effect of Z on X and the causal effect of X on Y
  - Mathematically we need to split effect of on Z on Y into effect of Z on X and X on Y

Formal Definition of an Instrumental Variable

Model:

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

- We call a variable Z a a valid **instrumental variable** if the following two conditions hold:

1. **Relevance:** $Cov(X, Z) \neq 0$
   - An arrow from Z to X in the pictures
2. **Exogeneity:** $Cov(U, Z) = 0$
   - No arrow from Z to Y or Z to U in the picture

Key Assumption #1 of Instrumental Variables

- **Relevance:** $Cov(X, Z) \neq 0$
- This assumption just means that X and Z are correlated
- We observe both X and Z, so can easily test this assumption by regressing:

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

- If $\beta_1 \neq 0$ in regression, we say instrument is relevant

Key Assumptions #2 of Instrumental Variables

- **Exogeneity:** $Cov(U, Z) = 0$
- This assumption means that Z and U cannot be correlated
- We do not observe U, so we **cannot** test this assumption
- In general, we need to "defend" this assumption by telling a story about why Z and U are unlikely to be correlated
    - Discuss this **a lot** later, but basically want to say that Z randomly assigns different X to individual

Best Defense of Exogeneity IV Assumption: Randomized Experiment

- Back to our Project STAR class size example:

$$score_i = \beta_0 + \beta_1 CS_i + u_i$$

- where:
  - $score_i$: Test score of student i
  - CS_i: Class size of student i
- Suppose that we use a coin flip that sends kids that get a "head" to a small class and kids getting a "tails" to big class
  - This is our randomized experiment!
- Let's call the coin flip our instrument Z (where $Z_i = 1$ if heads, $Z_i = 0$ if tails)

Best Defense of Exogeneity IV Assumption: Randomized Experiment

$$score_i = \beta_0 + \beta_1 CS_i + u_i$$

- Is Z (our coin flip) a good instrument?
- **Relevance:** $Cov(CS, Z) \neq 0$? Yes, if $Z_i = 1$ kid gets small class, if $Z_i = 0$ kid gets big class
  - So the regression $CS_i = \beta_0 + \beta_1 Z_i + u_i$ will estimate that $\beta_1 < 0$
- **Exogeneity:** $Cov(U, Z) \neq 0$? Untestable – so need "storytime"
  - Story: Exogeneity holds because coin flip is random and does not depend on any student or parent characteristics that would affect test scores. Therefore, there is nothing related to Z (besides X) that is also related to test scores, so U and Z must be uncorrelated

Randomized Experiment as an IV

$$score_i = \beta_0 + \beta_1 CS_i + u_i$$

- So a randomized experiment can be treated as an IV
  - Big difference: can test randomized experiment somewhat by checking balance of covariates, while IVs often cannot be tested
  - For that reason, usually differentiate between IVs and randomized experiments
- Later we will see different Zs. Good way to form your "story" of whether they are good: think of whether they are "mimicking" a randomized experiment