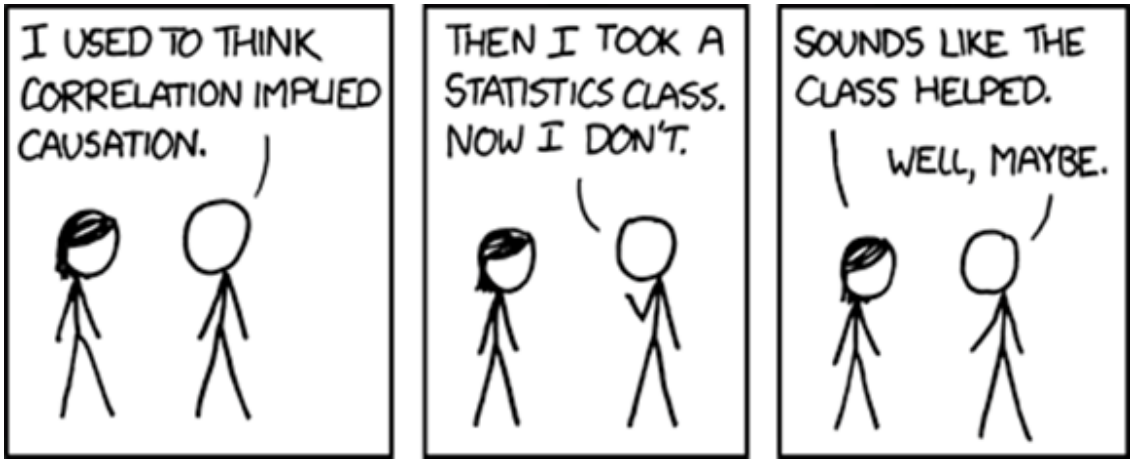
Chapter 11
Identifying Causation



Source: HTTP://XKCD.COM/552/

Learning Objectives

- Explain how correlation differs from causation in regression models
- Learn the three sources of the endogeneity problem and how they cause assumption CR5 to fail
- Learn about some solutions to the endogeneity problem

Why Causality

- Does joining a union or getting more education raise workers' earnings?
- Do immigrants reduce wages and employment for native workers?
- Do good roads make economies grow faster?
- Does economic growth make civil wars less likely?
- Do smaller class sizes help kids learn more?
- Do welfare payments to poor women make their kids better educated and well-fed?
- What makes some nations rich and others poor?

How to tell correlation from causation

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

- Correlation means that if you tell me *X*, I can make a prediction of *Y*.
  - The population doesn't change
- Causation means that if you ***change*** *X* to a different value, then I expect *Y* to change.
  - The population changes because there was a "treatment"

A New Assumption

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

CR5: *The values of the explanatory variable are exogenous, or given (there are no errors in the X-direction).*

- If CR5 holds, then β_1 is the causal effect of *X* on *Y*
- If CR5 fails, then we have an ***endogeneity problem***

Three Sources of "Endogeneity"

1. Measurement error in *X*.

2. *X* and *Y* determined jointly.

3. Omitted *X* variable.

Some economists use the word "endogeneity" only for jointly determined variables, but modern econometrics uses the word for any of the three settings.

1. Measurement Error

Regression you want: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

You observe *X* with error: $\tilde{X}_i = X_i + u_i$

Regression you run: $Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{X}_i + \tilde{\varepsilon}_i$

What's in the error? (Hint: substitute for X_i in the original equation)

Bias from Measurement Error

$$\tilde{\varepsilon}_i = \beta_1 \left(X_i - \tilde{X}_i \right) + \varepsilon_i = -\beta_1 u_i + \varepsilon_i$$

- So the error now has the original error, plus the error in X times β_1
- OLS estimate of β_1 usually biased towards zero

- Example: Janitor randomly messes up your experiment in the night

2. *X* and *Y* determined jointly

- Supply and demand

$$Q_i = \beta_0 + \beta_1 P_i + \beta_2 XD_i + \varepsilon_{Di}$$
$$Q_i = \alpha_0 + \alpha_1 P_i + \alpha_2 XS_i + \varepsilon_{Si}$$

Are you estimating the supply equation or the demand equation?

Solve for P_i and you'll find BOTH errors in it!

3. Endogeneity (Omitted Variables)

- Data: random sample of 30-39 year olds in the United States
- X_1 =1 if *i* has a college degree and X_i = 0 otherwise
- Y_i = log earnings last year

$$Y_i = 10.3 + 0.25X_i + e_i$$

- This means that college graduates earned 25% more than non-graduates
- **If the non-graduates had gone to college, would they have earned 25% more?**

Endogeneity = Omitted Variables

Regression you run: $Y_i = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^*$

Schooling

Earnings

But people with higher ability get more schooling:

("Model" of schooling) $X_{1i} = \alpha_0 + \alpha_1 Z_{2i} + u_i$

Regression you should run:

Ability

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 Z_{2i} + \varepsilon_i$$

Coefficients β_1^* and β_1 are the same *only* if X_1 and Z_2 are uncorrelated or $\beta_2 = 0$

What's the idealized "thought experiment?"

The Magic of Fixed Effects Models with Panel Data

- Multiple observations on same people over time (say, 2 years: t = 1, 2):

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 Z_{2i} + \varepsilon_{it}$$

Ability doesn't change over time (SO no "t" subscript)

- *Changes* in earnings:

$$Y_{i2} - Y_{i1} = (\beta_0 - \beta_0) + \beta_1(X_{1i2} - X_{1i1}) + \beta_2(Z_{1i} - Z_{1i}) + \varepsilon_{i2} - \varepsilon_{i1}$$

$$Y_{i2} - Y_{i1} = \beta_1(X_{1i2} - X_{1i1}) + \varepsilon_i^*$$

- So the missing ability variable disappears…PROBLEM SOLVED
  - But people's schooling has to change over time

What We Learned

- **Correlation** means that if you tell me *X*, I can make a prediction of *Y*.
- **Causation** means that if you *change X* to a different value, then I expect *Y* to change.
- The three sources of "endogeneity" are (i) measurement error, (ii) simultaneity, and (iii) omitted variables.
- **Measurement error** in *X* variables usually (but not always) leads to coefficient estimates that are smaller than they should be (biased toward zero). Proxy variables can help reduce measurement error bias.
- **Simultaneity** means that *X* and *Y* cause each other.
- Fixed-effects estimation can mitigate the omitted variables problem in panel data (but only for time-invariant omitted variables)