**Vanderbilt University Political Science - Stats I**
**Fall 2024 - Prof. Jim Bisbee**

# 18 Lecture 18

## 18.1 Confounds, revisited

[Go over "identify the potential confound."]

## 18.2 Omitted variable bias

- We've just explored several different ways that one might go about controlling for a variable. We are about to go into detail on one of the simplest (and perhaps least satisfying) way to do this: including an additive term with the potential confound, Z, in the linear model.
- Although this technique may seem overly simple, it can still provide us with unbiased estimates of the *ceteris paribus* relationship between X and Y if certain assumptions hold. To see this, let's first analyze what happens when we don't control for Z:
- Assume that the true model is

$$y = \beta_0 + \beta_1 x + \beta_2 z + u$$

- Notice that we're making a big assumption here about z: no interaction between x and z, and z enters into the DGP in a linear fashion.
- Because this model is properly specified, $\nu$ is an error term that does not covary with either $x$ or $z$ conditional on the other variable: i.e., $cov\,(u|x,z) = cov\,(u|z,x) = 0$.
- But let's say instead we regress y only on x, falsely assuming that the model is

$$y = \beta_0 + \beta_1 x + \nu,$$

- and thus incorrectly assuming that $cov\,(\nu, x) = 0$.

- then what we are really doing is moving $\beta_2 z$ to the error term, $\nu$ :

$$\begin{aligned} y &= \beta_0 + \beta_1 x + (\beta_2 z + u), \\ \text{where } \nu &= (\beta_2 z + u). \end{aligned}$$

- You'll recall that in the bivariate case that our estimator is

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{cov\,(x, y)}{var\,(x)}$$

(here writing $\nu$ instead of $u$):

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \overline{x})\,\nu_i}{SST_x},$$

- Here rely on the assumption that the covariance of x and $\nu$ is zero to make the final term dissappear, and thus say that $E\left(\widehat{\beta}_1\right) = \beta_1$. But now consider

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \overline{x})\,(\beta_2 z_i + u_i)}{SST_x}.$$

- Taking expectations, we now have

$$\begin{aligned} E\left(\widehat{\beta}_1\right) &= E\,(\beta_1) + E\left[\frac{\sum (x_i - \overline{x})\,(\beta_2 z_i + u_i)}{SST_x}\right] \\ &= \beta_1 + \frac{E\,(\sum x_i \beta_2 z_i + x_i u_i - \overline{x}\beta_2 z_i - \overline{x} u_i)}{SST_x} \end{aligned}$$

- Now we do two things. We (1) assume the z's are fixed (just as we do the x's in the bivariate case) and (2) we invoke the (correct) assumption that $E\,(u|x, z) = 0$. Now we can write:

$$\begin{aligned} &= \beta_1 + \frac{\sum x_i \beta_2 z_i - \overline{x}\beta_2 z_i}{SST_x} \text{or more helpfully,} \\ E\left(\widehat{\beta}_1\right) &= \beta_1 + \beta_2\left[\frac{\sum z_i\,(x_i - \overline{x})}{SST_x}\right]. \end{aligned}$$

- With a little manipulation, we see that

$$\frac{\sum z_i\left(x_i-\overline{x}\right)}{SST_x}=\frac{\sum z_i x_i-\sum z_i\overline{x}}{\sum\left(x_i-\overline{x}\right)^2}=\frac{\sum z_i x_i-n\overline{z}\overline{x}}{\sum\left(x_i-\overline{x}\right)^2}=\frac{S_{xz}}{S_{xx}}=\frac{cov(x,z)}{var(x)}$$

- And so it turns out that $z_i\frac{\sum(x_i-\overline{x})}{SST_x}=\frac{cov(x,z)}{var(x)}$, which is the slope coefficient we would obtain if we simply regressed z on x! So quite simply, we can write:

$$E\left(\widehat{\beta}_1\right)=\beta_1+\beta_2\frac{cov(x,z)}{var(x)},$$

- and thus

$$BIAS\left(\widehat{\beta}_1\right)=E\left(\widehat{\beta}_1\right)-\beta_1=\beta_2\frac{cov(x,z)}{var(x)}.$$

- What if we wanted to say something about the sign of the bias? Well, note that $sign\left[\frac{cov(x,z)}{var(x)}\right]=sign\left[cov(x,z)\right]$. So if we omit $z$ from our equation, we can now say that sign of $\widehat{\beta}_1$'s bias is

$$sign\left[cov(x,z)\times\beta_2\right]$$

- What does this mean in practice? Consider a regression in which you model feelings toward Barack Obama as a function of Democratic Party identification. You omit a dummy variable for whether an individual is African-American. In what direction is your estimate of $\beta_1$ almost assuredly biased?

- That is, you assume the model is

$$\begin{aligned}\text{ObamaFT} &= \beta_0+\beta_1\left(\text{DEM}\right)+\nu,\text{ when the true model is}\\ \text{ObamaFT} &= \beta_0+\beta_1\left(\text{DEM}\right)+\beta_2\left(\text{BLACK}\right)+u.\end{aligned}$$

- Well, we're pretty sure that $\beta_2>0$ and $cov\left(\text{DEM, BLACK}\right)>0$.

- So our estimate of $\beta_1$ will have a bias greater than zero. A.k.a., it is "biased upward,": we will overestimate the effect of Democratic Party identification because we are not accounting for African-American racial identity.

- What happens if $cov(x,z) = 0$? What happens if $\beta_2 = 0$?
  - That's right: as we've said before, when a variable is omitted, TWO problems must be present in order for it to cause bias:
    1. it is correlated with one or more $x$'s in your model.
    2. its partial effect on $y$ is not zero.
  - Why, then, do we love randomly assigning individuals to $x$? Because by construction, $cov(x,z)$ (for any omitted $z$ you can think of) is zero, making $\widehat{\beta}_1$unbiased.
- This is a nice simple example, but it gets more complicated in a multivariate context. You'll see this shortly.
  - [If the class asks: that's because the term $\beta_2 \left[\frac{\sum z_i(x_i - \bar{x})}{SST_x}\right]$ becomes $\beta_2 \left[\frac{1}{N}\left(X'X\right)^{-1}\left(X'z\right)\right]$, which takes into account the extent to which the omitted variable $(z)$ is collinear with all the included $x$'s in the model. In practice, the sign of this bias is hard to consider in such a back-of-the-envelope fashion.]
- Take-home point: if you leave out a variable that is BOTH correlated with included $x$'s and has a separate effect on $y$, your estimates will suffer from omitted variable bias.
- If this omitted variable enters into the true DGP in an additive linear fashion, we can obtain unbiased estimates of $\beta_1$ and $\beta_2$–that is, the *ceteris paribus* relationships of $y$ and $x$, and $y$ and $z$, respectively–by moving to multiple regression. But to do that, we need a little matrix algebra.

## 18.3 Revisiting matrix algebra

- Here, go over:
  - Matrix algebra handout I, pp. 1-3;
  - Handout IV (entire)