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

17 Lecture 17

Although controlling for a variable by adding Z as an additive term in a multiple regression seems overly simple, it can still provide us with unbiased estimates of the ceteris paribus relationship between X and Y.

- – To see this, let's first analyze what happens when we don't control for Z:
- Assume that the true model is

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

  where u is an error term such that $cor\,(u|x,z) = 0$. Regressing y on x1 and x2 will yield unbiased, consistent estimates of β.

- Notice that we're making a big assumption here: no interaction between x and z, and z enters into the DGP in a linear fashion.
- But if instead we regress y only on x1, obtaining the equation

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

  then what we are really doing is moving $\beta_2 x_2$ to the error term, ν :

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

- You'll recall that in the bivariate case,

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

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

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

- Taking expectations, we now have

$$\begin{aligned} E\left(\widehat{\beta}_1\right) &= E(\beta_1) + E\left[\frac{\sum (x_i - \overline{x})(\beta_2 z_i + \nu)}{SST_{x_1}}\right] \\ &= \beta_1 + \frac{\sum (x_{1i} - \overline{x}) E\left[(\beta_2 z_i + \nu)\right]}{SST_x} \\ &= \beta_1 + \beta_2 \left[z_i \frac{\sum (x_i - \overline{x})}{SST_x}\right]. \end{aligned}$$

- 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 regressed z on 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 x2 from our equation, we can now say that its sign 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 β_1 almost assuredly biased?

- What happens if $cov(x,z) = 0$? What happens if $\beta_2 = 0$?

  - That's right: 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 that next time.
  - [That's because the term $\beta_2 \left[x_2 \frac{\sum(x_{1i}-\overline{x_1})}{SST_{x_1}}\right]$ becomes $\beta_2 \left[\left(\frac{1}{N}X'X\right)^{-1}\left(\frac{1}{N}X'x_2\right)\right]$, which takes into account the extent to which the omitted variable (x_2) 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.