Simple derivation of omitted variables bias

EDS 222
Tamma Carleton

Omitted variables bias is a common violation of the exogeneity assumption of Ordinary Least Squares (OLS), and causes estimated regression coefficients to be biased relative to true population parameters. Omitted variables bias arises when there exists a variable that you are not including in your regression but that satisfies the following two conditions:

1. The omitted variable is correlated with your dependent variable of interest
2. The omitted variable is correlated with at least one of your independent variables

Note that if only one of these conditions is met, you do not have a problem. To see how this bias arises mathematically, suppose the following relationship represents the true population relationship between y and x_1 and x_2:

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

But suppose you only are really interested in x_1, and therefore you only include x_1 in your regression, ignoring x_2. What goes wrong?

First, note that condition #1 above holds as long as $\beta_2 > 0$, since the true population model tells us that a one unit change in x_2 causes a β_2 unit change in y. If the second condition also holds, we can write x_2 as a function of x_1:

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

If you do not include x_2 in your regression, its effect on y is subsumed in your error term, i.e. variation in y that is not explained by your model:

$$y = \beta_0 + \beta_1 x_1 + \nu,$$

where $\nu = \beta_2 x_2 + \varepsilon$.

We can substitute our expression for x_2 into this expression and rearrange terms to see that:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \nu \\ &= \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + e) + \varepsilon \\ &= \beta_0 + \beta_2\delta_0 + (\beta_1 + \beta_2\delta_1)x_1 + \beta_2 e + \varepsilon \end{aligned}$$

When we regress y only on x_1, ignoring x_2, we therefore obtain:

$$y = \underbrace{\beta_0 + \beta_2\delta_0}_{intercept} + \underbrace{(\beta_1 + \beta_2\delta_1)}_{slope} x_1 + \eta,$$

where η is mean zero because we assume that both e and ε are mean zero.

What's the problem? Our estimated intercept is now $\beta_0 + \beta_2\delta_0$ and our estimated slope is now $\beta_1 + \beta_2\delta_1$, both of which are biased estimators of the true β_0 and β_1 that we are after.

Note that these expressions help you think through which direction your bias is likely to go in practice. If β_2 and δ_1 are both positive, meaning that y and x_2 are positively related as well as x_2 and x_1, your slope coefficient will be biased upward when you omit x_2 (and therefore your estimated slope coefficient should fall when you add x_2 into the regression). In contrast, if either β_2 or δ_1 is negative, your slope coefficient will be biased downward when you omit x_2, and adding x_2 into your regression should increase your estimated slope coefficient.