Omitted Variable Bias: The Simple Case

Suppose that we omit a variable that actually belongs in the true (or population) model.

This is often called the problem of excluding a relevant variable or under-specifying the model.

This problem generally causes the OLS estimators to be biased.

Deriving the bias caused by omitting an important variable is an example of misspecification analysis.

Let us begin assuming that the true population model is

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

and that this model satisfies Assumptions MLR.1–MLR.4.

Primary interest: β_1, the partial effect of x_1 on y.

Example: y is log of hourly wage, x_1 is education, and x_2 is a measure of innate ability. To get an unbiased estimator of β_1, we should run a regression of y on x_1 and x_2 (which gives unbiased estimators of β_0, β_1 and β_2).

However, due to our ignorance or data unavailability, we estimate the model by excluding x_2.

In other words, we perform a simple regression of y on x_1 only, obtaining the equation

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

We use the symbol "$\tilde{}$" rather than "$\hat{}$" to emphasize that $\tilde{\beta}_1$ comes from an underspecified model.

We can derive the algebraic relationship

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the slope estimators (if we could have them) from the multiple regression

$$y_i \quad \text{on} \quad x_{i1}, x_{i2} \qquad i = 1, \ldots, n,$$

and $\tilde{\delta}$ is the slope from the simple regression

$$x_{i2} \quad \text{on} \quad x_{i1} \qquad i = 1, \ldots, n.$$

Because $\tilde{\delta}$ depends only on the independent variables in the sample, we treat it as fixed (nonrandom) when computing $E(\tilde{\delta})$.

Bias size

It is known that $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased for β_1 and β_2. Therefore,

$$\begin{aligned} E(\tilde{\beta}_1) &= E(\hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}) \\ &= E(\hat{\beta}_1) + E(\hat{\beta}_2)\tilde{\delta} = \beta_1 + \beta_2\tilde{\delta} \end{aligned}$$

which implies that the bias in $\tilde{\beta}_1$ is

$$\text{Bias}(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1 = \beta_2\tilde{\delta}.$$

Because the bias in this case arises from omitting the explanatory variable x_2, the term on the right-hand side of the above equation $(\beta_2\tilde{\delta})$ is often called the omitted variable bias.

It is easy to see that $\text{Bias}(\tilde{\beta}_1) = 0$ when

1. $\beta_2 = 0$
   The omitted variable x_2 is not in the "true" model.

2. $\tilde{\delta} = 0$
   Recall that $\tilde{\delta}$ is the slope from the simple regression

$$x_{i2} \quad \text{on} \quad x_{i1} \qquad i = 1, \ldots, n,$$

   which is directly related to the correlation between x_1 and x_2. Therefore, when x_1 and x_2 are uncorrelated, omitting x_2 does NOT lead to biased estimate of β_1, regardless of the value of β_2.

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Wage example

More ability $\Rightarrow$ higher productivity $\Rightarrow$ higher wages $\Rightarrow$ $\beta_2 > 0$ in

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{abil} + u,$$

Conjecture: educ and abil are positively correlated
On average, individuals with more innate ability choose higher levels of education.

Consequence: OLS estimates from

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + u,$$

are **on average** too large.