Omitted variable bias

The purpose of this derivation is to understand how omitted variable bias affects our estimate of a parameter of interest under conditions we can control. The derivation relies on the tower property of expectations[1] or the rule of iterated expectations, i.e., $E[y] = E[E[y|x]]$.

An econometrician wants to estimate the relationship between an outcome y and an explanatory variable x_1. Falsely, they assume the relationship

$$y_i = \gamma_0 + \gamma_1 x_{1i} + v_i,$$

which is *omitting* a second variable x_{2i}. The true model of the world, or, equivalently, the actual data generating process is as follow

$$\begin{aligned}
&\text{(population regression function)} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \quad E[u_i|x_1, x_2] = 0 && (1)\\
&\text{(ancillary regression function)} \quad x_{2i} = \delta_0 + \delta_1 x_{1i} + \varepsilon_i, \qquad E[\varepsilon_i|x_1] = 0. && (2)
\end{aligned}$$

Notice that it follows immediately from the assumptions in (2) that

$$E[x_{2i}|x_1] = E[\delta_0 + \delta_1 x_{1i} + \varepsilon_i|x_1] = \delta_0 + \delta_1 x_{1i} + \cancelto{0}{E[\varepsilon_i|x_1]} = \delta_0 + \delta_1 x_{1i}. \tag{3}$$

Using OLS, the econometrician estimates the slope parameter of their single linear regression as

$$\hat{\gamma}_0 = \frac{\sum_i^n (x_{1i} - \bar{x})(y_{1i} - \bar{y})}{\sum_i (x_{1i} - \bar{x})^2}, \tag{4}$$

where a bar indicates an average, e.g., $\bar{x} = \frac{1}{n}\sum_i^n x_{1i}$. Before, progressing with the derivations, it is useful to realize that a sum of demeaned observations is zero, step by step

$$\sum_i^n (x_{1i} - \bar{x}) = \sum_i^n x_{1i} - \sum_i^n x_{1i}\bar{x} = \sum_i^n x_{1i} - n\bar{x} = \sum_i^n x_{1i} - n\frac{1}{n}\sum_i^n x_{1i} = \sum_i^n x_{1i} - \sum_i^n x_{1i} = 0. \tag{5}$$

Therefore, the sum of the product of two demeaned variables can be rewritten as the product of a demeaned variable and a raw variable, e.g.,

$$\begin{aligned}
\sum_i^n (x_{1i} - \bar{x})(y_{1i} - \bar{y}) &= \sum_i^n (x_{1i} - \bar{x})y_{1i} - \sum_i^n (x_{1i} - \bar{x})\bar{y} = \sum_i^n (x_{1i} - \bar{x})y_{1i} - \bar{y}\cancelto{0}{\sum_i^n (x_{1i} - \bar{x})} \\
&= \sum_i^n (x_{1i} - \bar{x})y_{1i}. && (6)
\end{aligned}$$

We can, therefore, rewrite (4), substitute in the true value of y_i from (1), expand the expression and factor out constants

$$\begin{aligned}
\hat{\gamma}_0 &= \frac{\sum_i^n (x_{1i} - \bar{x})(y_{1i} - \bar{y})}{\sum_i (x_{1i} - \bar{x})^2} = \frac{\sum_i^n (x_{1i} - \bar{x})y_{1i}}{\sum_i (x_{1i} - \bar{x})^2} \\
&= \frac{\sum_i^n (x_{1i} - \bar{x})(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i)}{\sum_i (x_{1i} - \bar{x})^2} \\
&= \beta_0 \frac{\cancelto{0}{\sum_i^n (x_{1i} - \bar{x})}}{\sum_i (x_{1i} - \bar{x})^2} + \beta_1 \frac{\sum_i^n (x_{1i} - \bar{x})x_{1i}}{\sum_i (x_{1i} - \bar{x})^2} + \beta_2 \frac{\sum_i^n (x_{1i} - \bar{x})x_{2i}}{\sum_i (x_{1i} - \bar{x})^2} + \frac{\sum_i^n (x_{1i} - \bar{x})u_i}{\sum_i (x_{1i} - \bar{x})^2}, \\
&= \beta_1 \cancelto{1}{\frac{\sum_i^n (x_{1i} - \bar{x})^2}{\sum_i (x_{1i} - \bar{x})^2}} + \beta_2 \frac{\sum_i^n (x_{1i} - \bar{x})x_{2i}}{\sum_i (x_{1i} - \bar{x})^2} + \frac{\sum_i^n (x_{1i} - \bar{x})u_i}{\sum_i (x_{1i} - \bar{x})^2} = \beta_1 + \beta_2 \frac{\sum_i^n (x_{1i} - \bar{x})x_{2i}}{\sum_i (x_{1i} - \bar{x})^2} + \frac{\sum_i^n (x_{1i} - \bar{x})u_i}{\sum_i (x_{1i} - \bar{x})^2}
\end{aligned}$$

[1] Link to Wikipedia article.

where the simplifications follow from (5) and (6). For $\hat{\gamma}_0$ to be unbiased it must equal β_1 in expectation, however, taking the unconditional expectation and using the tower property (or rule of iterated expectations), i.e., $E[y] = E[E[y|x]]$, on the econometricians estimate we find that

$$\begin{aligned}
E[\hat{\gamma}_0] &= E\left[\beta_1 + \beta_2 \frac{\sum_i^n (x_{1i} - \bar{x}) x_{2i}}{\sum_i (x_{1i} - \bar{x})^2} + \frac{\sum_i^n (x_{1i} - \bar{x}) u_i}{\sum_i (x_{1i} - \bar{x})^2}\right] \\
&= \beta_1 + E\left[E\left[\beta_2 \frac{\sum_i^n (x_{1i} - \bar{x}) x_{2i}}{\sum_i (x_{1i} - \bar{x})^2} + \frac{\sum_i^n (x_{1i} - \bar{x}) u_i}{\sum_i (x_{1i} - \bar{x})^2} \middle| x_1, x_2\right]\right] \\
&= \beta_1 + E\left[\beta_2 \frac{\sum_i^n (x_{1i} - \bar{x}) x_{2i}}{\sum_i (x_{1i} - \bar{x})^2} + \frac{\sum_i^n (x_{1i} - \bar{x}) \cancelto{0}{E\left[u_i | x_1, x_2\right]}}{\sum_i (x_{1i} - \bar{x})^2}\right] \\
&= \beta_1 + \beta_2 E\left[\frac{\sum_i^n (x_{1i} - \bar{x}) x_{2i}}{\sum_i (x_{1i} - \bar{x})^2}\right],
\end{aligned}$$

and taking one of two paths, we either apply the tower property again, but only condition on x_1, substitute in the expectation from (3) and make simplifications based on (5) and (6) again

$$\begin{aligned}
E[\hat{\gamma}_0] &= \beta_1 + \beta_2 E\left[E\left[\frac{\sum_i^n (x_{1i} - \bar{x}) x_{2i}}{\sum_i (x_{1i} - \bar{x})^2} \middle| x_1\right]\right] = \beta_1 + \beta_2 E\left[\frac{\sum_i^n (x_{1i} - \bar{x}) E\left[x_{2i} | x_1\right]}{\sum_i (x_{1i} - \bar{x})^2}\right] \\
&= \beta_1 + \beta_2 \left(\delta_0 \cancelto{0}{\frac{\sum_i^n (x_{1i} - \bar{x})}{\sum_i (x_{1i} - \bar{x})^2}} + \delta_1 \cancelto{1}{\frac{\sum_i^n (x_{1i} - \bar{x}) x_{1i}}{\sum_i (x_{1i} - \bar{x})^2}}\right) = \beta_1 + \beta_2 \delta_1
\end{aligned}$$

or we recognize that

$$\frac{\sum_i^n (x_{1i} - \bar{x}) x_{2i}}{\sum_i (x_{1i} - \bar{x})^2} = \hat{\delta}_1,$$

and infer that $E[\hat{\delta}_1] = \delta_1$, based on (2), i.e., the fact that true model of x_2 and the zero-conditional mean assumption depends only on x_1. We find that the bias in the in the econometricians estimate due to the omitted variable is

$$\text{Bias}[\hat{\gamma}_1] = E[\hat{\gamma}_1 - \beta_1] = E[\hat{\gamma}_1] - \beta_1 = \beta_1 + \beta_2 \delta_1 - \beta_1 = \beta_2 \delta_1.$$