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: \( \beta_1 \), 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 \( \beta_1 \), we should run a regression of \( y \) on \( x_1 \) and \( x_2 \) (which gives unbiased estimators of \( \beta_0 \), \( \beta_1 \) and \( \beta_2 \)).

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 “~” rather than “^” 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

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

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

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

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 $\beta_1$ and $\beta_2$. Therefore,

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

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
   
   $$x_{i2} \text{ on } x_{i1} \quad i = 1, \ldots, n,$$
   
   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 $\beta_1$, regardless of the value of $\beta_2$.

<table>
<thead>
<tr>
<th>$\beta_2$</th>
<th>Corr($x_1, x_2$) $&gt; 0$</th>
<th>Corr($x_1, x_2$) $&lt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_2 &gt; 0$</td>
<td>Positive bias</td>
<td>Negative bias</td>
</tr>
<tr>
<td>$\beta_2 &lt; 0$</td>
<td>Negative bias</td>
<td>Positive bias</td>
</tr>
</tbody>
</table>
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.