Omitted Variables Bias
What happens when we omit an important variable?

Need to conjecture regarding the relationship between the omitted variable and included $x$ and $y$ variables

Upward bias:

- Estimate is on average higher than the true parameter: $E(\hat{\beta}|x) > \beta$

Downward bias:

- Estimate is on average lower than the true parameter: $E(\hat{\beta}|x) < \beta$

Let us derive rigorously when these cases occur.
Suppose that a population has the following relationship:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \]

However, we forget about \( x_2 \) and estimate:

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 \]

Our equation is *misspecified*.

What is the relationship between the true \( \beta_1 \) and \( \hat{\beta}_1 \)?

How do we quantify the bias?
Multivariate Regression
Omitted variable bias

- Population relationship between \( x_1 \) and \( x_2 \):
  \[
x_2 = \delta_0 + \delta_1 x_1 + \epsilon
\]
- Plug \( x_2 \) into the population equation:
  \[
y = \beta_0 + \beta_1 x_1 + \beta_2 (\delta_0 + \delta_1 x_1 + \epsilon) + u
\]
- Expand:
  \[
y = \beta_0 + \beta_1 x_1 + \beta_2 \delta_0 + \beta_2 \delta_1 x_1 + \beta_2 \epsilon + u
\]
- Group terms:
  \[
y = \underbrace{\beta_0 + \beta_2 \delta_0}_{\hat{\beta}_0} + \underbrace{(\beta_1 + \beta_2 \delta_1)}_{\hat{\beta}_1} x_1 + \underbrace{\tilde{u}}_{\hat{\beta}_2 \epsilon + u}
\]
- \( \hat{\beta}_1 \) is an estimate of \( \beta_1 + \beta_2 \delta_1 \), not \( \beta_1 \).
- Simple formula for the bias:
  \[
  \text{Bias} = \hat{\beta}_1 - \beta_1 = \beta_1 + \beta_2 \delta_1 - \beta_1 = \beta_2 \delta_1
  \]
To quantify the bias, we need the sign of $\beta_2$ and $\delta_1$.

If $x_2$ and $x_1$ are positively correlated, $\delta_1 > 0$. If not, then $\delta_1 < 0$.

If $x_2$ has a positive effect on $y$, $\beta_2 > 0$. If not, then $\beta_2 < 0$.

Put them together:

<table>
<thead>
<tr>
<th>$x_2$ has a positive effect on $y$</th>
<th>Corr$(x_1,x_2) &gt; 0$</th>
<th>Corr$(x_1,x_2) &lt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$(\delta_1 &gt; 0)$</td>
<td>$(\delta_1 &lt; 0)$</td>
</tr>
<tr>
<td>$x_2$ has a positive effect on $y$</td>
<td>positive bias</td>
<td>negative bias</td>
</tr>
<tr>
<td></td>
<td>$(\beta_2 \delta_1 &gt; 0)$</td>
<td>$(\beta_2 \delta_1 &lt; 0)$</td>
</tr>
<tr>
<td>$x_2$ has a negative effect on $y$</td>
<td>negative bias</td>
<td>positive bias</td>
</tr>
<tr>
<td></td>
<td>$(\beta_2 \delta_1 &lt; 0)$</td>
<td>$(\beta_2 \delta_1 &gt; 0)$</td>
</tr>
</tbody>
</table>
Example: Effect of class attendance on grades

Population follows:

\[ final = \beta_0 + \beta_1 \text{attend} + \beta_2 \text{study} + u \]

We instead forget about \textit{study} and estimate:

\[ \widehat{final} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{attend} \]

Suppose we estimate \( \widehat{\beta}_1 > 0 \), and conclude that attendance increases your grade (\( \beta_1 > 0 \)). Is this right?

- Positive correlation between \textit{study} and \textit{final}
- Positive correlation between \textit{study} and \textit{attend}
- \( \widehat{\beta}_1 \) suffers from an upward bias: \( \beta_1 < \widehat{\beta}_1 \)
Intuition

- \( \hat{\beta}_1 > 0 \) suggests that higher attendance improves your grade
- However, students who attend class often tend to study more
- Thus, attend may actually be accounting for the effects of studying, and not attendance.

Overall, given \( \beta_1 < \hat{\beta}_1 \), the result \( \hat{\beta}_1 > 0 \) is insufficient to guarantee that \( \beta_1 > 0 \).
Example: Effect of drugs on crime

Population follows:

\[ crime = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{drugs} + u \]

We instead forget about drugs and estimate:

\[ \hat{crime} = \hat{\beta}_0 + \hat{\beta}_1 \text{educ} \]

Suppose we estimate \( \hat{\beta}_1 < 0 \), and conclude education reduces your likelihood of committing a crime (\( \beta_1 < 0 \))

Positive correlation between drugs and crime

Negative correlation between drugs and educ

\( \hat{\beta}_1 \) suffers from an downward bias: \( \hat{\beta}_1 < \beta_1 \)
Multivariate Regression
Omitted variable bias - Examples

- **Intuition**
  - $\hat{\beta}_1 < 0$ suggests that education reduces your likelihood of committing a crime
  - However, people who go to school are less likely to abuse drugs
  - Thus, educ may actually be accounting for the propensity of drug use, not the effects of education
  - Overall, given $\hat{\beta}_1 < \beta_1$, the result $\hat{\beta}_1 < 0$ is *insufficient* to guarantee that $\beta_1 < 0$. 
Example: Effect of graduate education on wages

Population follows:

\[ \log(wage) = \beta_0 + \beta_1 \text{geduc} + \beta_2 \text{Exper} + u \]

We instead forget about \text{Exper} and estimate:

\[ \hat{\log(wage)} = \hat{\beta}_0 + \hat{\beta}_1 \log(\text{geduc}) \]

Suppose we estimate \( \hat{\beta}_1 > 0 \), and conclude that graduate education increases your wage (\( \beta_1 > 0 \))

Positive correlation between \text{Exper} and \( \log(wage) \)

Negative correlation between \text{Exper} and \( \text{geduc} \) (by construction)

\( \hat{\beta}_1 \) suffers from an downward bias: \( \hat{\beta}_1 < \beta_1 \)
Multivariate Regression
Omitted variable bias - Examples

- **Intuition**

  - $\hat{\beta}_1 > 0$ suggests that graduate education of some sort increases your wage.
  
  - However, people who pursue graduate education have lower levels of experience.
  
  - Thus, people with no graduate education may earn relatively high wages since they have lots of experience.

- Overall, given $\hat{\beta}_1 < \beta_1$, the result $\hat{\beta}_1 > 0$ is *sufficient* to guarantee that $\beta_1 > 0$. 