1 Omitted Variable Bias: Part I

Remember that a key assumption needed to get an unbiased estimate of β_1 in the simple linear regression is that $\mathbb{E}[u|x] = 0$. If this assumption does not hold then we can't expect our estimate $\hat{\beta}_1$ to be close to the true value β_1. We call this problem omitted variable bias. That is, due to us not including a key variable in the model, we have that $\mathbb{E}[\hat{\beta}_1] \neq \beta_1$. The motivation of multiple regression is therefore to take this key variable out of the error term by including it in our estimation.

2 Omitted Variable Bias: Part II

The formula for omitted variable bias can be a little confusing, so to start we'll go through a few things much more slowly. Remember those SLR1-5 assumptions we talked about last time? Prof. Buck stated in lecture that if SLR1-4 hold for a given model, then our estimates of the $\hat{\beta}$ will be unbiased. First we're going to take a closer look at what's going wrong once we start thinking about omitted variables.

SLR4 fails because of an omitted variable: $\mathbb{E}[u|X] \neq 0$

The Baseline: SLR.1-4 hold, and our estimates are unbiased

Population Model:

$$y = \beta_0 + \beta_1 x + u$$

Sample Regression:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

What's the OLS formula for $\hat{\beta}_1$?

$$\begin{aligned}
\hat{\beta}_1 &= \frac{Cov(x_i, y_i)}{Var(x_i)} \\
&= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x}) x_i} \quad \text{(See Appendix to these notes)}
\end{aligned}$$

We can use what we know about the population model, plug $y = \beta_0 + \beta_1 x + u$ into our formula for $\hat{\beta}_1$ and simplify:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum_i (x_i - \bar{x}) x_i} \\
&= \frac{\beta_0 \sum_i (x_i - \bar{x}) + \beta_1 \sum_i (x_i - \bar{x}) x_i + \sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x}) x_i} \\
&= \beta_1 + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x}) x_i}
\end{aligned}$$

Now, remember that $\hat{\beta}_1$ is a random variable, so that it has an expected value:

$$\mathbb{E}\left[\hat{\beta}_1\right] = \mathbb{E}\left[\beta_1 + \frac{\sum_i (x_i - \bar{x})u_i}{\sum_i (x_i - \bar{x})x_i}\right] = \beta_1 + \mathbb{E}\left[\frac{\sum_i (x_i - \bar{x})u_i}{\sum_i (x_i - \bar{x})x_i}\right] = \beta_1$$

Aha! So under assumptions SLR.1-4, on average our estimates of $\hat{\beta}_1$ will be equal to the true population parameter β_1 that we were after the whole time.

Reality Check: SLR.4 fails, $\mathbb{E}[u|X] \neq 0$, and our estimates are biased

Population Model:

Sample Regression:

What's the OLS formula for $\hat{\alpha}_1$?

$$\begin{aligned}\hat{\alpha}_1 = \frac{Cov(x_i, y_i)}{Var(x_i)} &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ &= \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x}) x_i}\end{aligned}$$

We can use what we know about the population model, plug y into our formula for $\hat{\alpha}_1$ and simplify:

$$\begin{aligned}\hat{\alpha}_1 &= \frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \beta_2 z_i + u_i)}{\sum_i (x_i - \bar{x}) x_i} \\ &= \frac{\beta_0 \sum_i (x_i - \bar{x}) + \beta_1 \sum_i (x_i - \bar{x}) x_i + \beta_2 \sum_i (x_i - \bar{x}) z_i + \sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x}) x_i} \\ &= \beta_1 + \beta_2 \frac{\sum_i (x_i - \bar{x}) z_i}{\sum_i (x_i - \bar{x}) x_i} + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x}) x_i}\end{aligned}$$

There's an extra term! The second term $\beta_2 \frac{\sum_i (x_i - \bar{x}) z_i}{\sum_i (x_i - \bar{x}) x_i}$ is a result of our omission of the variable z that affects y. When SLR.1-4 hold, on average our regression estimates will be close to the true parameters. But here, SLR.1-4 do not hold! If we take the expectation of $\hat{\alpha}_1$:

$$\begin{aligned}\mathbb{E}[\hat{\alpha}_1] &= \mathbb{E}\left[\beta_1 + \beta_2 \frac{\sum_i (x_i - \bar{x}) z_i}{\sum_i (x_i - \bar{x}) x_i} + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x}) x_i}\right] \\ &= \beta_1 + \beta_2 \mathbb{E}\left[\frac{\sum_i (x_i - \bar{x}) z_i}{\sum_i (x_i - \bar{x}) x_i}\right] + \mathbb{E}\left[\frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x}) x_i}\right] \\ &= \beta_1 + \beta_2 \rho_1\end{aligned}$$

If $\mathbb{E}[\hat{\alpha}_1] \neq \beta_1$ then we say $\hat{\alpha}_1$ is biased. What this means is that on average, our regression estimate is going to miss the true population parameter by _____.

3 Example: OVB in Action

In this section, I use the wage data (WAGE1.dta) from your textbook to demonstrate the evils of omitted variable bias and show you that the OVB formula works. Let's pretend (!) that this sample of 500

people is our whole population of interest, so that when we run our regressions, we are actually revealing the true parameters instead of just estimates. We're interested in the relationship between wages and gender, and our "omitted" variable will be tenure (how long the person has been at his/her job). Suppose our population model is:

$$\log(wage)_i = \beta_0 + \beta_1 female_i + \beta_2 tenure_i + u_i \tag{1}$$

First let's look at the correlations between our variables and see if we can't predict how omitting tenure will bias $\hat{\beta}_1$:

```
. corr lwage female tenure
             |    lwage   female   tenure
-------------+---------------------------
       lwage |   1.0000
      female |  -0.3737   1.0000
      tenure |   0.3255  -0.1979   1.0000
```

If we ran the regression:

$$\log(wage)_i = \alpha_0 + \alpha_1 female_i + e_i \tag{2}$$

...then the information above tells us that α_1 __ β_1. Let's see if we were right. Below is the Stata output from running regressions (1) and (2):

```
. reg lwage female tenure
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   67.64
       Model |  30.4831298     2  15.2415649           Prob > F      =  0.0000
    Residual |  117.846622   523  .225328148           R-squared     =  0.2055
-------------+------------------------------           Adj R-squared =  0.2025
       Total |  148.329751   525   .28253286           Root MSE      =  .47469
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3421323    .042267    -8.09   0.000    -.4251663   -.2590984
      tenure |   .0192648   .0029255     6.59   0.000     .0135176    .0250119
       _cons |   1.688842   .0343675    49.14   0.000     1.621326    1.756357
------------------------------------------------------------------------------
. reg lwage female
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =   85.04
       Model |  20.7120004     1  20.7120004           Prob > F      =  0.0000
    Residual |  127.617751   524  .243545326           R-squared     =  0.1396
-------------+------------------------------           Adj R-squared =  0.1380
       Total |  148.329751   525   .28253286           Root MSE      =   .4935
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3972175   .0430732    -9.22   0.000    -.4818349   -.3126001
       _cons |    1.81357   .0298136    60.83   0.000     1.755001    1.872139
------------------------------------------------------------------------------
```

Just to clarify, we "know" that β_1 =_____ and α_1 = _____.

This means that our BIAS is equal to: _____

There's one more parameter missing from our OVB formula. What regression do we have to run to find its value?

$$tenure = \rho_0 + \rho_1 female + v \tag{3}$$

The Stata output from this regression is below:

```
. reg tenure female
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =   21.36
       Model |  1073.26518     1  1073.26518           Prob > F      =  0.0000
    Residual |  26327.9839   524   50.244244           R-squared     =  0.0392
-------------+------------------------------           Adj R-squared =  0.0373
       Total |   27401.249   525  52.1928553           Root MSE      =  7.0883
------------------------------------------------------------------------------
      tenure |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.859373    .618672    -4.62   0.000    -4.074755   -1.643991
       _cons |   6.474453   .4282209    15.12   0.000     5.633212    7.315693
------------------------------------------------------------------------------
```

Just to clarify, our $\rho =$ _____

Now we can plug all of our parameters into the bias formula to check that it in fact gives us the bias from leaving out tenure from our wage regression:

$$\begin{aligned} \alpha_1 = \mathbb{E}[\hat{\alpha}_1] &= \\ \beta_1 + \beta_2\delta_1 & \\ &= \\ -.3421323 + (.0192648)(-2.859373) & \\ &= -0.397217549 \end{aligned}$$

4 OVB Intuition

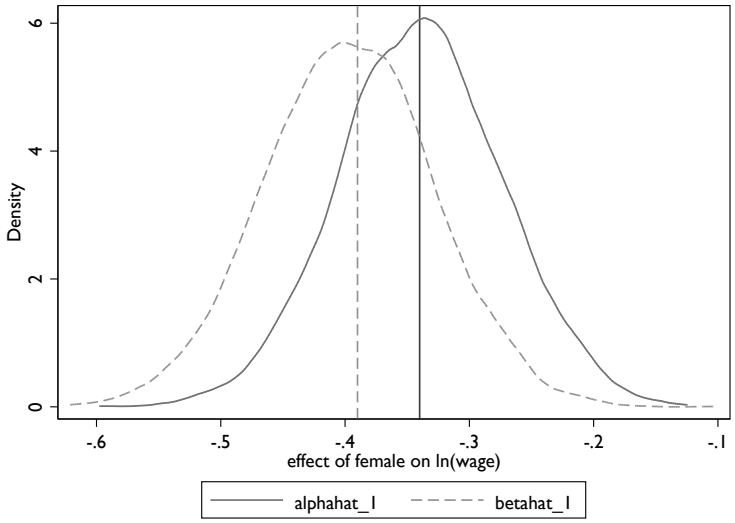
For further intuition on omitted variable bias, I like to think of an archer. When our MLR1-4 hold, the archer is aiming the arrow directly at the center of the target—if he/she misses, it's due to random fluctuations in the air that push the arrow around, or maybe imperfections in the arrow that send it a little off course. When MLR1-4 do not all hold, like when we have an omitted variable, the archer is no longer aiming at the center of the target. There are still puffs of air and feather imperfections that send the arrow off course, but the course wasn't even the right one to begin with! The arrow (which you should think of as our $\hat{\beta}$) misses the center of the target (which you should think of as our true β) systematically.

To demonstrate this, I did the following:

- Take a random sample of 150 people out of the 500 that are in WAGE1.dta
- Estimate $\hat{\beta}_1$ using OLS, controlling for *tenure* with these 150 people.
- Estimate $\hat{\alpha}_1$ using OLS (NOT controlling for *tenure*) with these 150 people.
- Repeat 6000 times.

At the end of all of the above, I end up with 6000 biased and 6000 unbiased estimates of $\hat{\beta}_1$. I plotted the kernel density of the biased estimates alongside that of the unbiased estimates. You can see how the biased distribution is shifted to the left indicating a downward bias!

Figure 1. Kernel densities for biased and unbiased estimates.



Take home practice problem: How to sign the bias

Traffic fatalities and primary seatbelt laws. Using data from Anderson (2008) for 49 US states, we can examine how primary seatbelt laws (an officer can pull you over just for not wearing your seatbelt) impact annual traffic fatalities. From the paper, I have data on the number of traffic fatalities in 2000, whether or not the state had a pimary seatbelt law in place, and the total population of the state. In 2000, just 35% of the 49 states had primary seatbelt laws (the rest had what's called a secondary seatbelt law). Suppose we run the following regression:

$$\widehat{fatalities} = \hat{\beta}_0 + \hat{\beta}_1 pop + \hat{\beta}_2 primary$$

1. Think of another variable or factor that you think affects traffic fatalities:

2. Is this factor positively or negatively correlated with $fatalities$? + or −
3. Is this factor positively or negatively correlated with $primary$? + or −
4. Omitting this factor from our regression will bias $\hat{\beta}_1$: UPWARD or DOWNWARD

Here are my results:

$$\widehat{fatalities} = 156.002 + 0.1232 pop + 17.258 primary$$

Whoa! According to our estimates, predicted fatalities increase with the implementation of a primary seatbelt law. Behavioral explanations aside[1], omitted variables are the likely culprits here. What are some variables that would induce an upward bias in $\hat{\beta}_2$?
I thought that weather might play a role in this puzzle. States with more "dangerous" weather will have more traffic fatalities and are also more likely to have a primary seatbelt law:

$$\widehat{fatalities} = 219.16 + 0.1204 pop - 78.092 primary + 68.963 precip - 579.91 snow$$

[1]By this I mean arguments like, "adding safety requirements results in people behaving more recklessly." While often valid—even in this particular case—we're going to keep it simple in this discussion.

(Clearly, even this specification with controls for weather has some issues: an additional inch of snow per year decreases predicted fatalities by 579.91 lives?)

5 Confidence Intervals

The simulation that was shown in section demonstrates something pretty profound: even after designing a random sample, collecting the data, figuring out the population model, and running regressions, **there's still a chance your estimates are very far from those of the population.** Each random sample yields a different estimate; if you have 100 random samples, you have 100 different values of $\hat{\beta}_1$. What can you do with them? Confidence intervals use the randomness of our sample estimates to say something useful about where the true population parameter actually is.

You can think of confidence intervals in two different ways:

1. We can think of a confidence interval as a bound for how wrong our sample estimate is. For example, if a political poll finds that a proposition will receive 53.2% of the vote, we come to very different conclusions if the "margin of error" is .5% or 5%.
2. Alternatively, we can think of a confidence interval as a measure of where the true, population value is likely to be. (The wording here is a little misleading, as you'll see in a bit.) For example, if the true average wage for US laborers is \$7, then it's unlikely that we'd find a confidence interval from our sample like [10,14].

The basics

We can think of a sample mean, $\bar{x}$ the same way we think about our $\hat{\beta}$s: these are both _____. We know even more about $\bar{x}$ from the Central Limit Theorem: For a random sample of a variable $\{x_1, ..., x_N\}$,

the Central Limit Theorem tells us that for very large samples (large N), the sample average $\bar{x} \sim N(\mu, \sigma^2_{\bar{x}})$. What this means: if I took 10,000 different random samples of laborers in the US and recorded their wage, I would end up with 10,000 different sample means $\{\bar{x}_1, \bar{x}_2, ..., \bar{x}_{10,000}\}$. If I plotted a histogram of all of these sample means, it would look very much like a normal distribution and the center of the distribution would be very close to the true average wage, μ, in the US. Because it's easy to get confused when we're talking about a random variable X and another random variable $\bar{x}$, which is the sample mean of X, here's a table to keep things straight:

Population	Sample
Mean of X: μ_X	Sample Mean: $\bar{x} = \frac{1}{n}\sum_i x_i$, and $\mathbb{E}[\bar{x}] = \mu_X$
Variance of x: $Var(x)$ or σ^2_x	Sample Variance of x: $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$, and $\mathbb{E}[s^2] = \sigma^2_x$
Variance of $\bar{x}$: $Var(\bar{x})$ or $\sigma^2_{\bar{x}}$	Sample Variance of $\bar{x}$: $s^2_{\bar{x}} = \frac{s^2}{n}$, and $\mathbb{E}\left[s^2_{\bar{x}}\right] = \sigma^2_{\bar{x}}$

Normal distributions are tricky to work with, and it's easier to *standardize* normally distributed variables so that they have a mean of 0 and a variance of 1. Remember our formula to find the expected value and variance of a transformed variable... If v is normally distributed with expected value $\mathbb{E}[v]$ and $Var(v) = \sigma^2_v$:

$$\mathbb{E}\left[\frac{v - \mathbb{E}[v]}{\sigma_v}\right] =$$

$$Var\left(\frac{v - \mathbb{E}[v]}{\sigma_v}\right) =$$

Since we're interested in the distribution of $\bar{x}$ (which is normal), we can standardize it just like above so that: $\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} \sim N(0,1)$

Now we can use what we know about the distribution of standard normal variables to help us say something meaningful about what the true population mean, μ_X might be:

- We know that for any standard normal variable v, $Pr(-1.96 < v < 1.96) = 95\%$
- We know that $\frac{\bar{x}-\mu_X}{\sigma_{\bar{x}}}$ is standard normal

But we're not *really* interested in the variable $\frac{\bar{x}-\mu_X}{\sigma_{\bar{x}}}$. The whole point of this is to learn more about μ_X! So we need to do some manipulation of this to isolate μ_X:

A standard normal distribution looks like this:

If we draw a number z from a standard normal distribution, then we know $Pr(1.96 < z < 1.96) =$ ____

Above we argued that $\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} \sim N(0,1)$ which means that the $Pr(-1.96 < \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} < 1.96) =$ ____

Just like in lecture, we can rearrange terms to see that $Pr(\bar{x} - 1.96\sigma_{\bar{x}} < \mu < \bar{x} + 1.96\sigma_{\bar{x}}) =$ ____

How to interpret a confidence interval

The most important thing to remember about a confidence interval is that the _____ is what's random, *not* the _____.

To make another metaphor out of an archaic sport, I like to think of confidence intervals in the context of a game of horseshoes:

When to use the Student t distribution:

As you could guess from the table, in practice we do not know what $\sigma_{\bar{x}}$ is, and we have to estimate it using sample data with: $s^2_{\bar{x}}$, which we call the standard error. Using the standard error (which is itself a random variable) changes the distribution of the sample mean a little bit, and we have to use a Student's t distribution instead of a Normal distribution:

$$\frac{\bar{x} - \mu}{\frac{s^2}{\sqrt{n}}} \sim t_{n-1}$$

The formula for a W% confidence interval is:

$$CI_W = \left[\bar{x} - c_W\left(\frac{s}{\sqrt{n}}\right), \bar{x} + c_W\left(\frac{s}{\sqrt{n}}\right)\right]$$

Where c_W is found by looking at the t-table for $n-1$ degrees of freedom.

Step 1. Determine the confidence level.

If we want to be 95% confident that our interval covers the true population parameter, then our confidence level is 0.95. Pretty straight forward.

Step 2. Compute your estimates of $\bar{x}$ and s.

Step 3. Find c from the t-table.

The value of c will depend on both the sample size (n) and the confidence level (always use 2-Tailed for confidence intervals):

- If our confidence level is 80% with a sample size of 10: $c_{80} =$
- If our confidence level is 95%, with a sample size of 1000: $c_{95} =$

Step 4. Plug everything into the formula and **interpret**.

Plug our values of $\bar{x}$, s and c into the given formula for a confidence interval. The trickiest step is the interpretation:

- Where does the randomness in a confidence interval come from?
- So when we construct a confidence interval, we interpret it by saying:

Example I took a random sample of 121 UCB students' heights in inches, and found that $\bar{x} = 65$ and $s^2 = 4$. Following the 4 steps above, I can find the 95% confidence interval for $\bar{x}$:

1. Confidence level was given: 95%
2. $\bar{x} = 65$ and $s^2 = 4$ were given.
3. From the t-table, $c = 1.987$
4. The 95% confidence interval is $[65 - 1.987 * \frac{\sqrt{4}}{\sqrt{121}}, 65 + 1.987 * \frac{\sqrt{4}}{\sqrt{121}}] = [64.64, 65.36]$. This interval has a 95% chance of covering the true average height of the population.

Practice

You have a random sample of housing prices in the Bay Area. After loading the data into Stata, you look at summary statistics for the prices you observed:

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       price |        88     293.546    102.7134        111        725
```

Find a 99% confidence interval for the true average housing price:

6 Appendix

Two facts used in the discussion of omitted variable bias:

$$\begin{aligned}\sum_i (x_i - \bar{x})(y_i - \bar{y}) &= \sum_i (x_i - \bar{x})y_i - \sum_i (x_i - \bar{x})\bar{y} \\ &= \sum_i (x_i - \bar{x})y_i - \bar{y}\sum_i (x_i - \bar{x}) \\ &= \sum_i (x_i - \bar{x})y_i - \bar{y}(0) = \sum_i (x_i - \bar{x})y_i\end{aligned}$$

Now replace every y_i in what's above to an x_i, and every $\bar{y}$ to an $\bar{x}$, and you can see that the same steps show $\sum_i (x_i - \bar{x})^2 = \sum_i (x_i - \bar{x})x_i$.