EEP/IAS 118
Andrew Dustan
Section Handout 4

1. Comparing assumptions for simple and multiple linear regression

The simple and multiple linear regression assumptions (SLR and MLR) are very similar. In fact, the simple assumptions are just special cases of the multiple assumptions. Here is a table comparing them:

	Simple	***Multiple***
S/MLR.1	Population model is linear in parameters: $y = \beta_0 + \beta_1 x + u$	Population model is linear in parameters: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$
S/MLR.2	$\{(x_i, y_i), i = 1, \dots, n\}$ is a random sample from the population	$\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i), i = 1, \dots, n\}$ is a random sample from the population
S/MLR.3	Not all sampled x values $\{x_i: i = 1, \dots, n\}$ are the same (x not constant)	No x is constant and there is no perfect collinearity among the x variables
S/MLR.4	No matter what the value of the observed variable (x), we expect the unobserved variable (u) to be zero: $E(u\|x) = 0$	No matter what the value of the observed variables $(x_1, x_2, \dots, x_k)$, we expect the unobserved variable (u) to be zero: $E(u\|x_1, x_2, \dots, x_k) = 0$
S/MLR.5	The "error term" (u) has the same variance for any value of the explanatory variable: $Var(u\|x) = \sigma^2$	The "error term" (u) has the same variance for any value of the explanatory variables: $Var(u\|x_1, x_2, \dots, x_k) = \sigma^2$

2. Interpreting results of multiple regression

Interpreting the results from multiple regression is not much different than doing the same with simple regression results. Still, it's important to know how to do it, and it's a good chance to practice.

Example 1 (from Wooldridge example 3.1):
Performing a regression of college GPA on high school GPA and ACT score, we get

$$\widehat{colGPA} = 1.29 + .453\ hsGPA + .0094\ ACT$$

How can we interpret the coefficient on *hsGPA*? There are a couple of different ways, one that assumes MLR.4 is met (no omitted variables bias) and one that doesn't require MLR.4:

Don't assume MLR.4	
Assume MLR.4	

Example 2 (from Wooldridge chapter 4):
Let's get some practice with logs. Here is the result of a regression of MLB baseball players' salaries on years in the league, games played per year, career batting average, average home runs per year, and average RBIs per year:

$$\log(\widehat{salary}) = 11.19 + .069\, years + .013\, gamesyr + .00098\, bavg + .014\, hrunsyr + .011\, rbisyr$$

Pick one of the estimated parameters and interpret it using a sentence with the word "predicted":

Explanatory variable	Interpretation

Example 3 (from Wooldridge exercise 3.4):
Now to look at salaries for a profession more attainable to most of us. Here is the result of a regression of median salary for new law school graduates on their LSAT score, median undergraduate GPA of the class, number of volumes in the law library, cost of attendance, and rank of the law school (1 being best). Each observation is one law school:

$$\log(\widehat{salary}) = 8.34 + .0047\, LSAT + .248\, GPA + .095\log(libvol) + .038\log(cost) - .0033\, rank$$

Interpret the coefficient on *libvol* (volumes in law library) using a sentence with the word "predicted":

The coefficient on *rank* seems pretty small. Does this mean that the rank of a law school doesn't matter a lot for graduates' salaries? Think in terms of the *partial* effect, holding the other explanatory variables constant.

3. Omitted variables bias

We formalized the problem of omitted variables bias (OVB) in lecture by supposing that a true population model was:

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

but that we specify the model incorrectly:

$$y = \widetilde{\beta_0} + \widetilde{\beta_1} x_1 \qquad\qquad + \tilde{u}$$

We then estimate this second (wrong) model using OLS. Why did we leave out x_2? Maybe we didn't realize it was important, but more likely, we didn't have data on x_2 so we couldn't include it. Why the tildes (squiggly lines) above the letters? We're basically acknowledging that if we could estimate each model using OLS, we would expect different estimates of the β's in each equation, so we shouldn't give them the exact same symbol. Note that in the second equation, $\beta_2 x_2$ is actually included in $\tilde{u}$ (there's no room for it elsewhere!).

Review your lecture notes for the rest, but in the end you get the result that if x_1 and x_2 are correlated, then leaving out x_2 will give us a biased (WRONG) estimate for the marginal effect of x_1 ($\widetilde{\beta_1}$).

The following exercise will try to give some intuition on how to think about OVB and how it is likely to affect regression results.

Steps for understanding potential OVB problems in a regression[1]:
To keep things simple, suppose you've estimated a regression of the following form:

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

Do the following, in order:

1. Think about which important variables might be missing from the regression. This can be any variable that affects the outcome variable (*y*) but that is not the same as the x_1 in your regression. Once you think of one (you always can), call that omitted variable x_2 and proceed to step 2.

2. Think of how x_1 and x_2 are related. Are they positively correlated, or negatively correlated? Using your answer, make the following table:

When x_1 is LOW	→	x_2 is (LOW) or (HIGH)
When x_1 is HIGH	→	x_2 is (LOW) or (HIGH)

3. Now think about how x_2 likely affects *y*. (You're making a guess about what β_2 is in the true population model.) Make a little table:

Ceteris paribus, when x_2 is LOW:	→	*y* is (LOW) or (HIGH)
Ceteris paribus, when x_2 is HIGH:	→	*y* is (LOW) or (HIGH)

4. Here's where you need to use your imagination. PRETEND that x_1 has NO effect on *y,* even if you know this is untrue. You're pretending the true β_1 is zero. Now use the above two tables to see how you expect *y* to differ between people with high and low values of x_1, **even if x_1 had no effect on *y***:

x_1 LOW	→	x_2 (LOW) or (HIGH)	→	*y* (LOW) or (HIGH)
x_1 HIGH	→	x_2 (LOW) or (HIGH)	→	*y* (LOW) or (HIGH)

5. Remember that we don't observe x_2, so let's modify the above table to reflect what we would actually observe (still assuming β_1 is zero).

x_1 LOW	→		→	*y* (LOW) or (HIGH)
x_1 HIGH	→		→	*y* (LOW) or (HIGH)

   Now we see the nature of the OVB problem. Even if we pretend $x_1$has no effect on *y*, we would still see some relationship between them in the data. This is due to the omitted variable.

6. Finally, formalize the relationship that we found in the table above. Pretending that the true β_1 is zero, what would our estimate $\hat{\beta}_1$ be due to this bias?
   If we found x_1 LOW→ *y* LOW, then they are positively related, so $\hat{\beta}_1 > 0$.
   If we found x_1 LOW→ *y* HIGH, then they are negatively related, so $\hat{\beta}_1 < 0$.

   But we had pretended the true effect was 0, so all we've done is find the sign of the OVB (we "signed the bias"). We are now free to stop pretending and allow $\beta_1 \neq 0$, with the sign of the OVB remaining what we found in our thought experiment. That is, if we found x_1 LOW→ *y* LOW, then $\hat{\beta}_1$ is too big ("biased upward"), and if we found x_1 LOW→ *y* HIGH, then $\hat{\beta}_1$ is too small ("biased downward").

This will become more intuitive with practice, of which we will do a lot.

[1] This discussion is based on Ben Crost, UC Berkeley ARE.