Econometrics
Chapter 6: Linear Regression with Multiple Regressors

Does the OLS assumption #1 hold? Or do we have Omitted Variable Bias?

Omitted Variable Bias

Lets look at the class size and test score example again.

We found that smaller class sizes were related to increased test scores.

Is this a causal effect? Maybe there is a problem.

1. In California (where the data came from) there are large immigrant populations and for these children English is not their native language.
2. And most of these types of children live in poorer districts.
3. And because of the way schools get tax funding from property tax, poorer districts tend to have larger classes.

$\rightarrow$ So the students in the larger classes are also more likely to not speak English as their native language which likely effects their test scores.

This is a problem if we care about causation.

Def: if our regressor is (1) *correlated* with a variable that is omitted from the regression (and so in the error term) and (2) this variable partly *determines* the outcome, then we say we have **omitted variable bias**.

This breaks our first assumption: $E[u_i|X_i] = 0$

A Note on Timing: The concern is with variables that meet the above, and occur before our X of interest. Sometimes this is not exactly clear and this is where one needs to think about what is the causal mechanism.

The OLS in this case has the limit of:

$\hat{\beta}_1 \rightarrow \beta_1 + \rho_{xu}\frac{\sigma_u}{\sigma_X}$

1. the bias does not go away even in large samples, and:
2. the size depends on how strongly X and u are correlated (ρ) and
3. the direction of the bias depends on the way they are related.

How do we fix this?

First lets look at breaking up the data according to how many non-native English speakers a group of schools have.

	Small Class		Large Class		Difference in Scores	
	Ave. Score	n	Ave. Score	n	Difference	t-stat (of diff)
All districts	657.4	238	650.0	182	7.4	4.04
% Non-native speakers						
$< 2\%$	664.5	76	665.4	27	-0.9	-0.30
1.9-8.8%	665.2	64	661.8	44	3.3	1.13
8.8-23%	654.9	54	649.7	50	5.2	1.72
$> 23\%$	636.7	44	634.8	61	1.9	0.68

So if we look at our basic (dummy) regression of 'small class' on test scores there is a pretty big effect (7.40).

But when we break it up, not only are all the effects much smaller but look not really significant. What happened?

Look at the composition of large/small classes with districts with less non-native speakers.

Less non-native speakers *also* have more small classes. And it is these classes that have the highest test scores.

So we can see that what we want to do is somehow 'control' for number of non-native speakers. We do this with a **multiple regression model**.

2nd Note on Timing: The structure of districts is what led to the class sizes, and the percent of English speakers is part of that structure. So in this sense omitting this variable is a problem.

But look at the opposite case: Should we control for the amount of time teachers spend with each student? Certainly time spent with student is correlated with class size and it is surely a determinant of test scores. But time-with-student happens after class size is determined, so it is one channel through with a class size affects test scores - we do not want to control for this. *We want this included.* Its a matter of the question you are asking.

Multiple Regression Model

This allows us to look at how some X_1 affects Y while holding other X's constant (for example % of non-native english speakers).

Say only have two X's, then our relationship can be modeled as:

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1$

This is our population model: the relationship between Y and the X's that holds on average in the population

β_0 is our intercept and other β's are our slope coefficients, or simply our coefficients on our X's.

Now our interpretation of β_1 is how changing X_1 changes Y *holding* X_2 *constant.*

Of course this relationship will not hold exactly for every person, so we add an error term:

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

Of course this is easily extended to any number of X's. So, how do we estimate these β's? - OLS.

The only thing that really changes (other than some refinement to our OLS assumptions) is the meaning of β_1.

Now, instead of $\beta_1 = \frac{\Delta Y}{\Delta X_1}$ its just $= \frac{\Delta Y}{\Delta X_1}$ holding all other Xs constant.

So you can think of it as a 'partial' effect.

OLS Estimator in Multiple Regression

Pick the $b_0, b_1, b_2 ..... b_k$ to estimate the $\beta_0, \beta_1, \beta_2 ..... \beta_k$ by minimizing the sum of the square errors of the predicted Y from the actual Y:

$\min_{b_0, b_1 ... b_k} \sum (Y_i - b_0 - b_1 X_{1i} - b_2 X2i - ..... - b_k X_{ki})^2$

Mathematical Details omitted

Test Score Example

Old $\beta_{STR} = -2.28$
New $\beta_{STR} = -1.1$; $\beta_{PctEL} = -0.65$

The new β_{STR} is smaller now because this reflects the effect when we hold % of non-native speakers constant.

The old effect really was picking up the effect of class size *and* % of non-native speakers

Measures of Fit in Multiple Regression

Standard Error of Regression (SER)

Estimate of the standard deviation of the error term (u), measure of spread of Y around the regression line.

$SER = s_{\hat{u}}; s^2_{\hat{u}} = \frac{1}{n-k-1} \sum \hat{u}_i^2 = \frac{SSR}{n-k-1}$

The difference between this one and the one for single regressor is the divisor is (n-k-1) instead of (n-2), this is because here we have degree of freedom correction for the estimated (k) slope plus 1 intercept coefficients.

The R^2

As before, the R^2 is the fraction of the sample variance of Y explained by the regression, or 1 minus the fraction of the variance of Y not explained by the regression.

$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$

$ESS = \sum(\hat{Y}_i - \overline{Y})^2$
$TSS = \sum(Y_i - \overline{Y})^2$
$SSR = \sum(u_i)^2$

But note that adding more regressors *always* increases your R^2.

Remember OLS minimizes sum of square errors, so adding another regressor either doesn't change it or must reduce it (because could have just made the coefficient on the new one zero and been just as good).

So we want to 'correct' our R^2 to sort penalize simply throwing in nonsense regressors.

Adjusted R^2

$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s^2_{\hat{u}}}{s^2_Y}$

Note the correction by the fraction $\frac{n-1}{n-k-1}$ which is always greater than 1 (and is embedded in $s^2_{\hat{u}}$) so the adjusted is always smaller than the unadjusted R^2

An interesting thing to note is that adjusted R^2 can be negative.

In many cases being too concerned about these measures of fit can be a trap. Don't just look at R^2 and think because it went up by adding a variable that the new model is better than the first.

OLS Assumptions in Multiple Regression

First three just extensions of original three, only the fourth is really new

#1. $E(u_i|X_{1i}, X_{2i}, ...X_{ki}) = 0$. This is the key assumption that makes OLS estimators unbiased.

#2. $(X_{1i}, X_{2i}, ..., X_{ki}, Y_i), i = 1, 2, ..., n$ are i.i.d.

#3. Large Outliers are unlikely.

#4. No Perfect Multicollinearity. The case where one regressor is a perfect linear function of another (or set of). If there is then the OLS estimators cannot be computed - its like dividing by zero in matrix setting.

Distribution of OLS Estimators in Multiple Regression

If our assumptions hold then:

1. The estimators are unbiased and constistent
2. In large samples the sampling distribution is well approximated by a normal distribution.
-So all the $\hat{\beta}s$ are jointly normal and each is also distributed normally.

Muliticollinearity

Dummy variable trap: Suppose you want to control for region with 'dummies'.

Say having three groups: North, central, and south. And you create a 0,1 variable for each: $N = \{0, 1\}, C = \{0, 1\}, S = \{0, 1\}$

If you include all three you will have perfect multicollinearity: N + C + S = 1 = intercept.

So what you need to do is leave one out (your 'base' group), and interpret the coefficients on the others as the difference between that group and the one left out.

Other ways to get perfect multicollinearity are simply making mistakes.

For example: Are any of your variables functions of each other?
Do you have a dummy variable for being male but all of your data is on males?

Imperfect Multicollinearity

This arises when two regressors are highly correlated.

The result is that the estimates of these coefficients will have high variances - they will not be precisely estimated (though still unbiased).