Notes 11: OLS Theorems

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

For a while we talked about the regression method. Then we talked about the linear model. There were many details, but the main takeaway is the following: the regression line is the best linear predictor of the expected output for a given value of the explanatory variables. The linear model assumes that the expected output is indeed linear. Then the best linear predictor will be the best predictor.

In this class, we will formalize this idea. We will tie the regression method and the linear model together. In other words, we will see how the OLS regression method relates to the linear model.

The question we are trying to answer is the following. Let the model be

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u,$$

where $\mathbb{E}[u|x_1, \ldots, x_k] = 0$. Suppose that we run an OLS regression of y onto $x_1, \ldots, x_k$, and we get a line:

$$y = a + b_1 x_1 + \cdots + b_k x_k.$$

Can we relate the a and $b_1, \ldots, b_k$ to the β's? Ideally we would like the OLS regression coefficients to be the same as the β's in the model. We would like to discover what the β's are. Can we use the OLS as a guessing method to discover the β's? That is, can we use a as a good guess of the value of β_0, b_1 as a good guess of β_1, and so on?

Remember what is an estimator? An estimator is a guessing method which uses the data. Well, the OLS regression uses the data, so the coefficients of the OLS regression are estimators of the β's in the regression. However, any other crazy combination of the data would also be an estimator. Why should we use the OLS coefficients as estimators of the β's? In other words, why should we use the OLS coefficients as a way to guess the value of the β's?

It all depends on what we want from an estimator. A good estimator has to:

- ____

This means that the estimator should be expected to guess the right value. It doesn't mean that it guesses the right value every time. It means that in average it will guess right.

- ____

  This means that the estimator shouldn't be guessing too far away from the true value.

It turns out that the two properties above have very clear mathematical equivalents. We explore them next. Before we move to understand the properties of the OLS estimators of the β's, we should give them a notation which is true to the custom in the profession. As you probably know from your Stat classes, usually we denote estimators with a hat. So, an estimator of β_0 is denoted $\hat{\beta}_0$.

In the rest of this course, unless I say something to the contrary, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the coefficients of the OLS regression. This means that $\hat{\beta}_0 = a$, $\hat{\beta}_1 = b_1$, $\ldots$, $\hat{\beta}_k = b_k$.

1 Does the OLS guess right?

The estimator should be expected to guess right. In other words, we want to check if:

In English: do we expect the OLS coefficients to be the same as the β's in the linear model? For this we need to be certain that the population and the data satisfy certain requirements.

Assumption 1 (population assumption): ____________________________

____

You are familiar with this assumption. It says that the world has to behave exactly as the model says. This can be quite a bit to ask from the world, and often it will not be true. For now (and for a while still) we will assume that this is indeed true.

The most important thing to observe about this assumption is that it is two conditions in one.

- ____
- ____

Assumption 2 (data assumption): ____________________________

This condition is entirely new. We never discussed it. Same as with the previous one, it is actually two conditions in one.

- ____
- ____

We shall discuss one at a time.

Random Sampling: The first condition is that the data was randomly collected. It means that the people that participated in the survey were collected at random from the population. This is not a trivial assumption. It often fails. We will discuss this condition in great detail in future classes.

No perfect multicollinearity: Multicollinearity is when the variables are linearly related in the sample. If you can write, for some $c_0, c_1, \ldots, c_k$, for all $i = 1, \dots, n$, then you have multicollinearity. So, if for example one of the variables is constant, say $x_{1i} = 3$ for all $i = 1, \dots, n$, then

If, for example, $x_{1i} = 5 + 2x_{2i}$, then

The point is that this cannot happen with the data we collected. You may ask: but what if the $x_1, \dots, x_k$ in the population are multicollinear, then the sample would be multicollinear as well. There would be nothing we could do to avoid it! The answer is that in the population, the $x_1, \dots, x_k$ cannot be multicollinear. Why?

____

____

This condition is a bit less problematic than the random sampling one. Perfect multicollinearity won't really cause much trouble, we will learn how to fix it easily. The problem is when we have almost multicollinearity (known as near multicollinearity). This is a harder to fix problem, and we will discuss it in depth in the future as well.

We finally arrive at the very anticipated result:

Theorem 1. ____

I want you to take a second to reflect about what this theorem is saying. It is saying that if the conditions above hold, then the OLS estimators (the coefficients of the regression line) guess exactly the right values in the linear model. We studied the interpretation of the coefficients, and how the coefficient of the variable of interest was the causal effect of that variable. What this theorem is saying is that if we run an OLS regression then the numbers we get are expected to be right. Are they always right? No. But at least they aren't consistently wrong. In average we are guessing right.

We have a technical term to describe the property "guessing right." We say that the OLS estimator is ____.

2 Does the OLS guess close?

We know that the OLS guesses right, but how close is it guessing? We will see that it depends on a number of things. How can we measure how close the OLS is guessing? We could look at the guess errors:

$$\hat{\beta}_j - \beta_j,$$

and see what we expect the errors to be. But wait, we just saw that

so that is no help. We could look at $\mathbb{E}[|\hat{\beta}_j - \beta_j|]$, but it turns out that it is hard to deal with this quantity mathematically. We do what we always do. We look at the squared errors

Ok, we need to find out what the variance of the OLS estimators $\hat{\beta}_j$ are. For this, we will need another assumption. You are a bit familiar with it already.

Assumption 3 (population assumption): ______________________________

____

____

This assumption is requiring one more thing from the population. We have a name for this condition, which we discussed in previous classes:

- ____

This condition is quite strong. However, it is not super important, and can be easily relaxed. For now, we will use this very strong condition to study the variance of the OLS estimators, because it makes everything easier. We will study what it really means, and what we can do when it fails in future classes as well.

Theorem 2. ____

____

____

____

Now we will study each component of the variance.

- σ^2: the variance of the unobservables. Of course, it would make sense that the more what we don't know varies, the less precise is our estimator.

- n: the sample size. Again, it makes sense that if we have more observations, we have more information about the population, and thus the estimator will be more precise.

- $\widehat{Var}(x_j)$: the sample variance of x_j. This one is a little more delicate. We spoke about it in the last class. It is easier to determine the inclination of a line if we have data that is horizontally spread out. This is directly related to the variation of x_j in the sample.

- $1 - R_j^2$: the multicollinearity measure. Remember that the R^2 of a regression of x_j on the other regressors is equal to one when all observation lie in a line. What is this? Multicollinearity! So, this is why we cannot have multicollinearity, otherwise the variance would be infinite. However, observe that "near-multicollinearity," which is when the R_j^2 is close to one, is also bad. It makes the estimate be very imprecise indeed.

Explaining the intuition behind it is hard. The idea you should have in your mind is that if we have perfect multicollinearity, then it is as if we had included two measures of the same thing. Perhaps they can be in a different unit (like Celsius and Fahrenheit), but they are basically the same. For example, suppose that $x_2 = x_1$. Then,

In English: we can never find out who β_1 and β_2 are. The best we can hope for is to find out $\beta_1 + \beta_2$.

Notice: ____

____

____