Recitation Four - Testing
Econometrics - Fall 2018
Regression

Nathaniel Mark

October 16, 2018

0.1 Omitted Variable Bias

First, think about the set up:
The True model is:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \gamma X2, i + \epsilon_i$$

with ϵ_i iid. Or in other words,

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

with

$$u_i = \gamma X_{2,i} + \epsilon_i$$

We run this model with only X_1. What happens? $\hat{\beta}_1$ is biased! Specifically, in this case with just one covariate and one omitted variable, we know that our estimator of β_1 takes the form:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2}$$

$$= \beta_1 + \frac{\sum(X_i - \bar{X})u_i}{\sum(X_i - \bar{X})^2}$$

And, by the law of large numbers,

$$\rightarrow \beta_1 + \frac{cov(X_i, u_i)}{var(X_{1,i})}$$

Plugging in our true model's definition of u_i and simplifying:

$$= \beta_1 + \frac{cov(X_i, \gamma X_{2,i} + \epsilon_i)}{var(X_{1,i})}$$

$$= \beta_1 + \frac{\gamma cov(X_i, X_{2,i}) + cov(X_i, \epsilon_i)}{var(X_{1,i})}$$

and since X_i and ϵ_i are independent,

$$= \beta_1 + \frac{\gamma cov(X_i, X_{2,i})}{var(X_{1,i})}$$

Therefore,

$$\hat{\beta}_1 \to \beta_1 + \frac{\gamma cov(X_i, X_{2,i})}{var(X_{1,i})}$$

So our estimator is NOT consistent, and its bias is a function of the effect of the omitted variable on the dependent variable (γ) and the covariance between the variable of interest ($X_{1,i}$) and the omitted variable.

Also note that the sign of the bias is equal to the "sign of the effect of the omitted variable on the dependent variable" times the "sign of the correlation between the relevant dependent variable and the omitted variable". This generalizes to cases with many variables.

0.2 F-tests

Once we have estimated a multi-variable regression:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}$$

A t-test tests the null hypothesis that **one** of these parameters are equal to a value. For example, we usually want to answer the question "can we statistically be certain that the effect of X_{ji} on Y_i is not zero?" Then, we use a t-test to test the null hypothesis that $\beta_j = 0$.

Often, we care about questions that involve more than one parameter. For example, we could ask "can we statistically be certain that the cumulative effect of X_{2i} and X_{3i} on Y_i is not zero?" Then, we would use the null hypothesis that $\beta_2 + \beta_3 = 0$. Why cant we just use a t-test?

A Valid t-test of this null hypothesis would be $\frac{(\hat{\beta}_2+\hat{\beta}_3)-0}{SE(\hat{\beta}_2+\hat{\beta}_3)} \sim^a N(0,1)$. The problem is that $SE(\hat{\beta}_2+\hat{\beta}_3) = \sqrt{var(\hat{\beta}_2)+var(\hat{\beta}_3)+2cov(\hat{\beta}_2,\hat{\beta}_3)}$. We can estimate all these, but it is rather complicated!

Second, what if we have more complex questions. What if you want to ask: Are the effect of X_{1i} on Y_i, the effect of X_{2i} on Y_i, and the effect of X_{3i} on Y_i all the same? Then, we would use the null hypothesis that $\beta_1 = \beta_2$ and $\beta_1 = \beta_3$. Now, using a t-test seems impossible. To test both of these hypotheses, we can use an F-test!

Outlines of an F-test:
1. Null hypothesis: 1 or more linear equations of β parameters.
2. Q - # of linear equations in the null hypothese (easiest way to find Q: count the number of equal signs you see in the null hypothesis.

(Do not need to know): The actual equation of the F-stat is a complicated matrix equation that you DO NOT NEED TO KNOW:

$$(R\hat{\beta}-r)'[R(\hat{V_n}/nR']^{-1}(R\hat{\beta}-r)/Q \sim^a F_{Q,n-k-1}$$

as $n \to \infty$.
where $H_0 : R\beta = r$, Q is the number of equation in the null hypothesis and k is the number of variables in the regression.

However, this equation gets a lot easier to handle in *special cases* which you should know (at least special case 1 and 3):

Special case One: When there is just one equation in the null hypothesis (e.g. $H_0 : \beta_1 = 0$), then

$$F-stat = (t-stat)^2$$

.

Special case Two: When there are two equations in the null hypothesis (e.g. $H_0 : \beta_1 = 0 \text{and} \beta_2 = 0$), then

$$F-stat = \frac{1}{2}\frac{(t-stat_1)^2+(t-stat_2)^2-2corr(t-stat_1,t-stat_2)}{1-corr(t-stat_1,t-stat_2)^2}$$

.

Special case Three: When Homoskedasticity holds,

$$F - stat = \frac{(SSR_R - SSR_u)/Q}{(SSR_R)/(n-k-1)} \approx$$

. after a little rearranging, this is equivalent to

$$F - stat = \frac{(R_U^2 - R_R^2)/Q}{(1-R_U^2)/(n-k-1)} \approx$$

Where U refers to the unrestricted regression (i.e the regression where the null hypothesis is not assumed) and R refers to the restricted regression (i.e the regression where the null hypothesis is assumed).

Once we have attained a valid F-stat, we can use it to test whether the null hypothesis is valid. That is, the estimator

F-stat $\sim$