

## Notes on F-tests

In the *Notes on Regression*, I review tests of hypotheses about regression coefficients using the confidence interval or t-test. This is a useful kind of test if we are interested in examining one coefficient at a time. But sometimes we are interested in testing hypotheses about more than one coefficient at a time. Let me consider a concrete example.

Suppose I estimate the following earnings equation for a sample of adult women. I have included as regressors years of school and dummy (0-1) variables for black and Latino:

$$\ln(Y) = b_0 + b_1 \text{SCHOOL} + b_2 \text{BLACK} + b_3 \text{LATINO}$$

Here are the results using Excel:

### Regression #1

#### SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.39731084
R Square	0.1578559
Adjusted R Square	0.14986087
Standard Error	0.56795969
Observations	320

#### ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	3	19.10717734	6.369059	19.74423	9.3348E-12
Residual	316	<b>101.9347143</b>	0.322578		
Total	319	121.0418916			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	8.87009126	0.188328217	47.09911	6.6E-145	8.49955534	9.240627173
SCHOOL	0.08924066	0.013320019	6.69974	9.58E-11	0.06303351	0.115447811
BLACK	-0.1486917	0.105200307	-1.41342	0.158518	-0.3556734	0.058289977
LATINO	-0.1583261	0.110814779	-1.42875	0.154065	-0.37635427	0.059702062

One hypothesis I might test is whether, controlling for years of school, black women earn more or less than white women. The hypothesis that there is no difference, once we control for SCHOOL and LATINO, is that the coefficient on BLACK is zero (that is, the hypothesis is  $\hat{\alpha}_2 = 0$ ). Examining my

regression results, I note that 0 is within the 95% confidence interval, so I DO NOT reject the hypothesis that the coefficient is zero. Thus these data suggest that there is *no significant difference* in earnings between black and white women, controlling for schooling. I could perform a similar hypothesis test on the coefficient for LATINO (it too is not significantly different from zero).

But suppose I wanted to test the following sort of hypothesis: *Neither BLACK nor LATINO has an effect on earnings— that is, both have zero effect. This hypothesis can be written as follows:  $\hat{\alpha}_2 = \hat{\alpha}_3 = 0$ .* You might sensibly think of testing this hypothesis using two separate t-tests on the two coefficients. But it turns out that this test is going to be too difficult to reject. Basically, it is possible to have each coefficient be insignificant when considered separately, but taken together still reject the hypothesis that BOTH are zero.

The appropriate test for a hypothesis about two or more coefficients is an F-test. The idea of this F-test can be understood in the following way. Suppose our hypothesis that neither BLACK nor LATINO affected earnings ( $\hat{\alpha}_2 = \hat{\alpha}_3 = 0$ ) were true. *Then if we dropped both variables from the regression and just used SCHOOL, it should fit the data nearly as well.* The sum of squared residuals (SSR) would go up a little, because this always happens when you drop a variable, but it wouldn't go up by much.

The F-statistic is just a formal statistical way of judging whether the SSR has changed enough to conclude that the coefficients on the variables we dropped had significant explanatory power.

Here's the regression when BLACK and LATINO are not included:

## Regression #2

### SUMMARY OUTPUT

<i>Regression Statistics</i>					
Multiple R		0.384423968			
R Square		0.147781787			
Adjusted R Square		0.145101856			
Standard Error		0.569547168			
Observations		320			

  

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	17.8877871	17.88779	55.14387	1.0367E-12
Residual	318	<b>103.1541045</b>	0.324384		
Total	319	121.0418916			

  

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	8.752211819	0.178014583	49.1657	1.2E-150	8.40197689	9.102447
SCHOOL	0.095654471	0.01288121	7.425892	1.04E-12	0.07031132	0.120998



In dropping the two variables, our SSR has gone from 101.93 in Regression #1 to 103.15 in Regression #2. Is this a significant change? We need to form the F-statistic, which is the following:

$$F = \frac{(SSR_2 - SSR_1)/m}{SSR_1/(n-k-1)}$$

where  $SSR_1$  is the SSR for the regression that includes the extra regressors,  $SSR_2$  is the SSR for the regression without those regressors,  $m$  is the number of regressors you have dropped (in this case,  $m = 2$ ),  $n$  is the number of observations, and  $k$  is the number of regressors in Regression #1 ( $k = 3$  in this case, because we have SCHOOL, BLACK, and LATINO). So for our hypothesis, the F-statistic is

$$F = \frac{(SSR_2 - SSR_1)/m}{SSR_1/(n-k-1)} = \frac{(103.15 - 101.93)/2}{101.93/(320-3-1)} = 1.89$$

Finally, we need to compare this value with the critical value on an F table. If our F is bigger than the critical value, we can reject the hypothesis. For our data sets, the relevant critical values for a test at the 95% confidence level are as follows:

m = 1	m = 2	m = 3	m = 4	m = 5
3.84	3.00	2.60	2.37	2.21

In our example, since  $m = 2$ , the appropriate critical value is 3.00. Our F-stat of 1.89 is below this critical value, so we DO NOT reject the null hypothesis that both coefficients are zero. That is, taken as a group, our variables for ethnicity in this regression do not have a significant effect on earnings.

OR: You can let Excel calculate a p-value for you. Suppose you have calculated the above F-stat using this formula and have it in a cell. Then all you need to do is enter the appropriate numbers in the following Excel formula: =FDIST(F, m, n-k-1). In our case, =FDIST(1.89, 2, 316) returns 0.153, which is well above the conventional significance threshold of 0.05, so again we DO NOT reject the null hypothesis that both coefficients are zero.