Econ 452 Section 7 - STATA

Connor Cole

November 12, 2015

Regression, Continued

T-Tests Continued

In our last section, we motivated t tests by assuming that errors were normally distributed, in which case (assuming that observations are iid) the following is true in exact finite sample distribution:

$$\frac{\hat{\beta}_j - \beta_j^{True}}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

Previously, we used this fact to test null hypotheses of the form $\beta_j^{True} = \gamma$, where γ is some value - for example 0. The logic was that since the above exact finite sample distribution holds if we have put in the 'true' value of β_j^{True}, we can compute the probability of observing the value we see of $\hat{\beta}_j$ under the assumption that the null hypothesis is true, or the p-value. Assuming our test statistic is sufficiently far out in the tails given some predetermined level of α, we would reject our null hypothesis. Keep in mind that this test involves the test of a single restriction, the null hypothesis claim that the value of $\beta_j^{True} = \gamma$. There are other null hypotheses that involve more than one parameter that in fact have a single restriction. For example, the null hypothesis claim that $\beta_j = \beta_i$ is a null hypothesis involving a single restriction, as is the null hypothesis claim that $\beta_j = \beta_i + 1$, or the null hypothesis claim that $\beta_j = \beta_i + \beta_q + 1$. In fact, the t-test works as a test of any single restriction null hypothesis. We can use the following fact, similar to the claim we see above, assuming that the null hypothesis is true:

$$\frac{Restriction^{True} - Restrction^{Observed}}{se(Restrction^{Observed})} \sim t_{n-k-1}$$

For example, if our null hypothesis is that $\beta_j = \beta_i + 1$, then our null hypothesis' restriction is $\beta_j - \beta_i - 1 = 0$, then assuming our null hypothesis is true:

$$\frac{Restriction^{True} - Restrction^{Observed}}{se(Restrction^{Observed})} \sim t_{n-k-1}$$

$$\frac{(\beta_j - \beta_i - 1) - (\hat{\beta}_j - \hat{\beta}_i - 1)}{se(\hat{\beta}_j - \hat{\beta}_i - 1)} \sim t_{n-k-1}$$

$$\frac{(\hat{\beta}_j - \hat{\beta}_i - 1)}{se(\hat{\beta}_j - \hat{\beta}_i - 1)} \sim t_{n-k-1}$$

$$\frac{(\hat{\beta}_j - \hat{\beta}_i - 1)}{se(\hat{\beta}_j - \hat{\beta}_i)} \sim t_{n-k-1}$$

$$\frac{(\hat{\beta}_j - \hat{\beta}_i - 1)}{\sqrt{V(\hat{\beta}_j) + V(\hat{\beta}_i) - 2cov(\hat{\beta}_j, \hat{\beta}_i)}} \sim t_{n-k-1}$$

The t test with a single restriction null hypothesis involving two parameters we discussed in class (and that you are more likely to see in applied settings), is the claim that for two parameters $\beta_j = \beta_i$. Then, the single restriction of the null hypothesis here is that $\beta_j - \beta_i = 0$. Using the framework we've discussed:

$$\frac{Restriction^{True} - Restrction^{Observed}}{se(Restrction^{Observed})} \sim t_{n-k-1}$$

$$\frac{(\beta_j - \beta_i) - (\hat{\beta}_j - \hat{\beta}_i)}{se(\hat{\beta}_j - \hat{\beta}_i)} \sim t_{n-k-1}$$

$$\frac{\hat{\beta}_j - \hat{\beta}_i}{\sqrt{V(\hat{\beta}_j + V(\hat{\beta}_i) - 2cov(\hat{\beta}_j, \hat{\beta}_i)}} \sim t_{n-k-1}$$

After deriving our test statistic and specifying the probability of a Type I error, α, then we can just use the same testing procedure as we've used before. If you need to review how testing works with t tests, look at the previous section.

F-Tests

We use t tests for null hypotheses that involve a single restriction on the values of variable coefficients. When a null hypothesis involves more than one restriction, then we use an F test. For example, if our null hypothesis is that $\beta_i = \beta_j = 0$, then in fact there are two restrictions embedded in this statement: $\beta_i = 0$ and $\beta_j = 0$. Our, if our null hypothesis is that $\beta_i - \beta_q = \beta_j = 0$, then again our null hypothesis has two restrictions: $\beta_i - \beta_q = 0$ and $\beta_j = 0$. An F test uses the fact that, assuming our restrictions in the null hypothesis are true, then:

$$T = \frac{\frac{SSR_{Restricted} - SSR_{Unrestricted}}{q}}{\frac{SSR_{Unrestricted}}{n-k-1}} \sim F_{q,n-k-1}$$

Where q is the number of restrictions applied, $SSR_{Unrestricted}$ is the sum of squared errors in the 'unrestricted' model where the q restrictions are not applied, $SSR_{Restricted}$ is the sum of squared errors in the 'restricted' model where the q restrictions are applied and the regression is re-estimated.

Using the fact that, mathematically, $SST = SSE + SSR$, then it is easy to derive that, assuming our restrictions in the null hypothesis are true, then:

$$\frac{\frac{R^2_{Unrestricted} - R^2_{Restricted}}{q}}{\frac{1-R^2_{Unrestricted}}{n-k-1}} \sim F_{q,n-k-1}$$

We can apply the F test for a series of restrictions in our null hypotheses after a regression by typing `test` and then putting in parentheses the restrictions we would like to test. For example, if we wanted to test the joint null hypothesis that $\beta_{VariableName2} = 3$, and $\beta_{VariableName2} = \beta_{VariableName3}$, we would type the line below. Note that even though three variable coefficients are included here, there are only two restrictions:

```
regress VariableName1 VariableName2 VariableName3 VariableName4

test (VariableName2-3 = 0) (VariableName2 - VariableName3 = 0)
```

Typing these commands, we get the following output:

STATA reports the two restrictions applied by our hypothesis test first, and then reports that the test statistic from the F test is 13807.55. Since we have applied two restrictions, and since $n - k - 1 = 3336$, this test statistic has an $F(2, 3336)$ distribution assuming the null hypothesis is true. The probability of observing a value larger than the test statistic we see is essentially 0.

Figure 1: F-Test Output

```
. test (VariableName2-3 = 0) (VariableName2 - VariableName3 = 0)

 ( 1)  VariableName2 = 3
 ( 2)  VariableName2 - VariableName3 = 0

       F(  2,  3336) =13807.55
            Prob > F =    0.0000
```

Let's spell out how to do an F test to test the hypothesis that $\beta_{VariableName2} = 0$, and $\beta_{VariableName3} = 0$ in the regression:

$$VariableName1 = \beta_1 + \beta_2 VariableName2 + \beta_3 VariableName3 + \beta_4 VariableName4 + \epsilon$$

1. **Define H_0.**
   $H_0 : \beta_{VariableName2} = 0$ and $\beta_{VariableName3} = 0$.

2. **Define H_1.**
   $H_1 : \beta_{VariableName2} \neq 0$ or $\beta_{VariableName3} \neq 0$.

3. **Define $\alpha =$ Probability of Type I Error.**
   As usual, we set $\alpha = 0.05$.

4. **Define the test statistic and test.**
   We apply an F test using the fact that we apply two restrictions in our null hypothesis and the number of observations is 3340 and the number of parameters estimated above is 4, then, assuming that our null hypothesis restrictions are true:

$$T = \frac{\frac{SSR_{Restricted} - SSR_{Unrestricted}}{q}}{\frac{SSR_{Unrestricted}}{n-k-1}} \sim F_{2,3336}$$

5. **Define a rejection region given the test.**
   We are in an F test, which is one sided and α has been set as 0.05, and the degrees of freedom are $3340 - 1 - 3 = 3336$. We need to find some value of C such that $P(T \geq C|H_0) = \alpha$, or the value C such that the probability of observing some test statistic T greater than C assuming the null is true is α. To find this value, we type in:

   ```
   display invF(2,3336, .95)
   ```

   We find that $C = 2.9984$

   Our rejection region is:

$$RR = (2.9984, \infty)$$

6. **Calculate and report the test statistic.** Using STATA's `test` command:

Figure 2: `Test` Command

```
. test (VariableName2=0) (VariableName3=0)

 ( 1)  VariableName2 = 0
 ( 2)  VariableName3 = 0

       F(  2,  3336) =  164.89
            Prob > F =    0.0000
```

We can derive this value by hand using the regressions as follows:

Figure 3: First Regression for F-Test

```
. regress VariableName1 VariableName2 VariableName3 VariableName4
```

Source	SS	df	MS	Number of obs	=	3,340
				F(3, 3336)	=	115.72
Model	2.5452e+11	3	8.4840e+10	Prob > F	=	0.0000
Residual	2.4458e+12	3,336	733164271	R-squared	=	0.0943
				Adj R-squared	=	0.0934
Total	2.7004e+12	3,339	808731655	Root MSE	=	27077

VariableName1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
VariableName2	-2209.933	429.1919	-5.15	0.000	-3051.439	-1368.427
VariableName3	.261637	.0169493	15.44	0.000	.2284049	.2948691
VariableName4	-500.3868	281.9362	-1.77	0.076	-1053.172	52.39867
_cons	20416.7	1225.731	16.66	0.000	18013.44	22819.96

Figure 4: Second Regression for F-Test

```
. regress VariableName1 VariableName4
```

Source	SS	df	MS	Number of obs	=	3,340
				F(1, 3338)	=	15.81
Model	1.2732e+10	1	1.2732e+10	Prob > F	=	0.0001
Residual	2.6876e+12	3,338	805159641	R-squared	=	0.0047
				Adj R-squared	=	0.0044
Total	2.7004e+12	3,339	808731655	Root MSE	=	28375

VariableName1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
VariableName4	-1164.924	292.9464	-3.98	0.000	-1739.296	-590.5509
_cons	32766.63	606.4718	54.03	0.000	31577.54	33955.73

$$T = \frac{\frac{SSR_{Restricted} - SSR_{Unrestricted}}{q}}{\frac{SSR_{Unrestricted}}{n-k-1}}$$

$$= \frac{\frac{2.6876e+12 - 2.4458e+12}{2}}{\frac{2.4458e+12}{3336}}$$

$$= 164.90408$$

$$T = \frac{\frac{R^2_{Unrestricted} - R^2_{Restricted}}{q}}{\frac{1 - R^2_{Unrestricted}}{n-k-1}}$$

$$= \frac{\frac{0.0934 - 0.0047}{2}}{\frac{1-0.0934}{3336}}$$

$$= 163.19391$$

7. **Do inference**. Under our choice of $\alpha = 0.05$, we see that the test statistic is very far into the rejection region. Therefore, we reject the null hypothesis at the 0.05 confidence level.

8. **Report p-values**. The p-value is $P(X > T|H_0)$, or the probability of observing an outcome as great or greater than the test statistic we observe given that the null is true. Note here that the p-value is reported under `Prob>F`. We see there that the p-value for this test listed in the table is essentially 0. We could derive this result by hand by typing in:

```
display Ftail(2,3336, 164.90408)
```

Where the `Ftail` function reports the probability of observing a value at or above 164.90408 in a F distribution with 2 and 3336 degrees of freedom.