FYI Handout #4: More on F tests

The ANOVA F test is used to test the hypothesis

(1) $$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } ... \text{ or } \beta_k \neq 0$$

in the multiple linear regression model

(2) $$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i, \qquad i = 1,2,...,n.$$

In words, the null hypothesis is that neither of the explanatory variables is linearly related to the response variable. The alternative hypothesis is that at least one of the explanatory variables is linearly related to the response variable. The alternative hypothesis is two-sided because it does not specify whether such association is positive or negative.

As we have seen, the test statistic is

(3) $$F = \frac{MSM}{MSE}.$$

If the null hypothesis is true, this statistic has an F distribution with *DFM* degrees of freedom in the numerator and *DFE* degrees of freedom in the denominator. *DFM*, the degrees of freedom of the model, are equal to the number explanatory variables in the regression (or the number of slope coefficients being estimated k). *DFE*, the degrees of freedom of the residual[1], are equal to the sample size (n) minus the total number of parameters being estimated (k slope parameters plus one intercept). Therefore, the distribution of F is $F(k, n - k - 1)$.

Comment 1: The above F statistic is closely related to the multiple coefficient of determination, defined as $R^2 = SSM / SST$. Given the definition of mean squares (sum of squares over degrees of freedom) and the fact that $SST = SSM + SSE$, we can rewrite the F statistic as

(4) $$F = \frac{SSM/k}{SSE/(n-k-1)} = \frac{SSM/kSST}{SSE/(n-k-1)SST} = \frac{R^2/k}{(1-R^2)/(n-k-1)}.$$

Therefore, high values of R^2 are likely to the associated to the rejection of the null hypothesis in (1).[2]

[1] E stands for "error", but it is more appropriate to call it residual, since the term error refers to the unobserved random variable ε_i in the regression model.

[2] Of course, all depends on the relationship between the sample size n and the number of explanatory variables k. The smaller is the sample size with respect to the number of explanatory variables, the *lower*

Comment 2: We can think the ANOVA *F* test as a test of imposing *k* linear *restrictions* on the parameters of the regression model (2). To see this, we can write the null hypothesis as

(1') $H_0 : \beta_1 = 0, \beta_2 = 0, ...,$ and $\beta_k = 0$.

It is clear that the null hypothesis contains *k* linear equations or restrictions on the values of the parameters.

There are many other tests of linear restrictions of interest in economics, beyond the ANOVA *F* test. For example, we are often interested in testing whether *some* (but not all) of the explanatory variables are linearly unrelated with the response variable. For example, we might be interested in testing whether the slope coefficients of the first *m* (*m* < *k*) explanatory variables are zero. In that case the hypothesis is

(5) $H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0$

$H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or} ... \text{or } \beta_m \neq 0$

Notice that the test is not concerned about the parameter values of the remaining $k - m$ explanatory variables ($\beta_{m+1}, \beta_{m+2}, ..., \beta_k$).

The computation of the test statistic involves running two different regressions: (1) a *restricted* regression, in which all the restrictions assumed to be true in the null hypothesis are imposed, and (2) an *unrestricted* regression with no restrictions. The restricted and unrestricted regressions corresponding to the hypothesis in (5) are

(6) Restricted: $y_i = \beta_) + \beta_m x_{mi} + \beta_{m+1} x_{m+1i} + ... + \beta_k x_{ki} + \varepsilon_i$

Unrestricted: $y_i = \beta_) + \beta_1 x_{1i} + ... + \beta_m x_{mi} + \beta_{m+1} x_{m+1,i} + ... + \beta_k x_{ki} + \varepsilon_i$

For each of these regressions we can obtain sums of squares, degrees of freedom, mean squares, and R^2s. Let's use the subscripts *U* for the unrestricted regression and *R* for the restricted regression. For example, SSE_U and DFE_U denote, respectively, the residual sum of squares and residual degrees of freedom of the unrestricted regression, while SSE_R and DFE_R are the counterparts of these concepts for the restricted regression.

The general test statistic for *m* linear restrictions to the parameters of the multiple linear regression model is

---

the value of the *F* statistic, irrespective of the R^2. For example, in a regression with $n = 10$, $k = 3$, and $R^2 = 0.7$, $F = 4.66$, with a *P*-value of 0.052. Even with this high R^2, the fails to reject the null hypothesis at the 5% significance level.

(7) $$F = \frac{(SSE_R - SSE_U)/(DFE_R - DFE_U)}{SSE_U / DFE_U}.$$

Notice that $SSE_R \geq SSE_U$ and $DFE_R > DFE_U$. The intuition for the first of these inequalities is the regression residuals should increase (or at least not decrease) if we restrict the values of the regression parameters. For example, consider a production function where output is a function of capital, labor, and energy. If we restrict the model so that energy does not enter in the equation (that is, if we constraint the coefficient of energy to be equal to zero), the differences between actual and estimated output (the residuals) are likely to increase. As for the second inequality, it is clear that the restricted model will have more degrees of freedom. This is because number of parameters to estimate in the restricted model (*m*) are less than in the unrestricted model (*k*); therefore $DFE_R = (n-(k-m)-1) > DFE_U = (n - k - 1)$

It follows that larger values of the *F* test statistic occur when the difference between SSE_R and SSE_U is greater, that is when the restrictions to the parameters greatly reduce the explanatory power of the model. To gain further intuition on the test statistic, divide numerator and denominator of (7) by the total sum of squares $TTS = \sum (y_i - \hat{y}_i)^2$. Noticing also that $DFE_R - DFE_U = (n-(k-m)-1)-(n-k-1) = m$, *F* can be written as

(8) $$F = \frac{(R_U^2 - R_R^2)/m}{(1-R_U^2)/(n-k-1)}.$$

This statistic follows the *F* distribution with *m* degrees of freedom in the numerator and $(n - k - 1)$ degrees of freedom in the denominator.

Notice that the ANOVA *F* test is a special case. When the restricted regression has only an intercept parameter, the restricted sum of squares residuals equals the total sum of squares, implying that $R_R^2 = 0$. Also, the total number of restrictions *m* equals the number of slope parameters *k*.

Implementing this test is simple: all it takes is to recover the R^2's and degrees o freedom of the restricted and unrestricted regressions, and then using Equation (8) to compute the *F* statistic. Then, the *P*-value is computed as the probability that a random variable having the $F(m, n - k - 1)$ distribution is greater than or equal to the value of the statistic.

Eviews allows you to implement tests of linear restrictions on the parameters of a multiple linear regression model. In the Equation window, select View→Coefficient Tests→Wald-Coefficient restrictions. Then enter the coefficient restrictions separated by commas, for example C(1)=0, C(2)=0 (where 1 and 2 are, respectively, the first and the second explanatory variables). The output shows both the *F* statistic along with a chi-square statistic for a similar test called the Wald test.

Examples:

1) The text contains one example of *F* tests for a group of regression coefficients in pp. 728-30.
2) An important application in economics is the Chow test for structural stability.

The Chow test asks whether or not a regression model is valid for all the observations in a sample. This test can be applied to time series data, to see whether a regression model changes in different periods, or to cross-section data, to see whether the model is different for different subpopulations. For example, consider a study of wage determination in a cross-section of workers. A researcher might suspect that the model is different for males and females. To apply the Chow test it is necessary, first, to state the restricted and unrestricted regressions. Assume that the sample is ordered so that the first n_1 observations correspond to females and the remaining n_2 correspond to males, and let $n = n_1 + n_2$ denote the total number of observations. If the model is the same for males and females, the regression could be specified as follows:

<u>Restricted Model</u>:

(9) $\log(WAGE)_i = \beta_) + \beta_1 ED_{1i} + \beta_2 EX_{2i} + \varepsilon_i, \quad i = 1,\ldots,n .$

On the other hand, if the model differs for males and females, the unrestricted model consists of two different regressions:

<u>Unrestricted Model</u>:

(10) Females: $\log(WAGE)_i = \beta_) + \beta_1 ED_{1i} + \beta_2 EX_{2i} + \varepsilon_{1i}, \quad i = 1,\ldots,n_1 .$

(11) Males: $\log(WAGE)_i = \gamma_) + \gamma_1 ED_{1i} + \gamma_2 EX_{2i} + \varepsilon_{2i}, \quad i = n_1 + 1,\ldots,n .$

The residual degrees of freedom are $DFE_R = n - 3$ for the restricted model and $DFE_U = n - 6$ for the unrestricted model (the latter has six parameters: $\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1,$ and γ_2). The residual sum of square of the restricted model SSE_R comes from the estimation of Equation (9). On the other hand, the residual some of squares of the unrestricted model equals the sum of the residual sums of squares of regressions (10) and (11): $SSE_U = SSE_F + SSE_M$. With all this information, a test of the hypothesis

(12) $H_0 : \beta_0 = \gamma_0, \beta_1 = \gamma_1, \text{and } \beta_2 = \gamma_3$

$H_1 : \beta_0 \neq \gamma_0 \text{ or } \beta_1 \neq \gamma_1 \text{ or } \beta_2 \neq \gamma_2$

can be performed by computing the *F* statistic

$$F = \frac{(SSE_R - SSE_U)/(DFE_R - DFE_U)}{SSE_U / DFE_U} = \frac{(R_U^2 - R_R^2)/3}{(1 - R_U^2)/(n-6)} .$$