The F Statistic

Suppose we regress y on three different regressors, w, x and z, and both slope coefficients for x and z have high enough p-values that we conclude each one is statistically insignificant. It is still possible, however, that they may be *jointly* significant, even if they are individually insignificant. In other words, we want to test simultaneously that

$$H_0 : \beta_x = \beta_z = 0,$$

$$H_A : \text{at least one of } \beta_x, \beta_z \neq 0.$$

Think of H_0 as being a *restriction* placed on β_x and β_z that we want to test.

The first thing to do is take the model where β_x and β_z are unrestricted (that is, a regression where x and z are included and thus their coefficients are estimated) and find its sum of squared residuals, call it RSS_{ur}. Then make the restrictions (by not even including them in the regression, which implicitly sets them equal to zero) and find that model's sum of squared residuals, call it RSS_r.

If β_x and β_z are jointly insignificant, i.e. if H_0 is true, then you would expect the difference between the two RSS terms to be small since the RSS represents unexplained variation in y. In other words, if β_x and β_z are jointly insignificant, then we shouldn't expect much difference in how well the model explain things whether they're both simultaneously included or not.

We just need to formalize what we mean by a "small" difference between the two. This is given by the F statistic,

$$F \equiv \frac{(RSS_r - RSS_{ur})/(k-g)}{RSS_{ur}/(n-k)} \sim F_{k-g,n-k}$$

where

- n is the number of observations;
- k is the number of parameters being estimated in the unrestricted model, in this case $k = 4$ because we estimate the intercept plus slope coefficients for w, x, and z;
- g is the number of parameters being estimated in the restricted model, in this case $g = 2$ because the restricted models omits w and z and hence only estimates the intercept and slope coefficient for w;
- $F_{k-g,n-k}$ is the F distribution with $k - g$ parameters included in the restriction and $n - k$ is the unrestricted degrees of freedom.

At the extreme end, we can also test whether *all* regressors are jointly significant by comparing it to a regression with *no* regressors. This yields the F statistic

$$F \equiv \frac{R^2/(k-1)}{(1-R^2)/(n-k)},$$

where R^2 is given in the unrestricted regression.

Example: Final 2016, Problem 3c

The question asks us to test whether the variables *radio*, *newspaper*, *tvbynews*, *region*1, and *region*2 are jointly significant. (Thus we already know that $k - g = 5$).

Unrestricted Model

The unrestricted model is the regression

$$sales = \beta_1 + \beta_2 tv + \beta_3 radio + \beta_4 newspaper + \beta_5 tvbynews + \beta_6 region1 + \beta_7 region2 + u.$$

Thus the unrestricted model estimates $k = 7$ parameters. The Stata output on the exam indicates that $RSS_{ur} = 518350292$.

Restricted Model

The restricted model, which omits the variables in question, is

$$sales = \beta_1 + \beta_2 tv + u,$$

and thus the restricted model estimates $g = 2$ parameters. The Stata output on the exam indicates that $RSS_r = 2102500000$.

F Statistic

There are $n = 200$ observations. Hence the F statistic is given by

$$F = \frac{(2102500000 - 518350292)/(7-2)}{518350292/(200-7)} \approx 117.97.$$

We are told that the test has a critical value of 2.261, and hence we reject the null because $117.97 > 2.261$.

Example: Final 2016, Problem 6c

Method 1

The R^2 statistic is the explanatory variation of y around its mean, i.e.

$$R^2 = \frac{540}{720} = 0.75.$$

The regression includes three estimates and thus $k = 3$. So the F statistic is

$$F = \frac{0.75/(3-1)}{(1-0.75)/(21-3)} = 27.$$

Method 2

Alternatively, consider the restricted model to be the one with no regressors, that is, where $H_0 : \beta_x = \beta_z = 0$. In this case, there is no explained sum of squares since there are no regressors doing any explaining! Thus $TSS_r = RSS_r = 720$ and $g = 1$ because only the intercept, incidentally the mean of y, is being estimated.

For the unrestricted case, i.e. the one where all of the regressors are used and estimated, we have $TSS_{ur} = ESS_{ur} + RSS_{ur}$ implies that $RSS_{ur} = 720 - 540 = 180$, and $k = 3$. Hence the F statistic is

$$F = \frac{(720-180)/(3-1)}{180/(21-3)} = 27.$$