F statistic for overall significance of a regression

Alessandro Bramucci

The F test

The F test is used to test whether a group of variables has no effect on the dependent variable. In this sense, the test allows to test if the parameters of a set (or at the limit all) the independent variables are **jointly significance**. Obviously it is the theory or intuition that tells us to operate such a test on a given group of variables. It is often the case that the F test is performed on all independent variables in a model. It is then said that the test is for the overall significance of the regression. In this exercise, to understand how the F test works in practice, we will replicate the F test provided by the regression function in R (as by any other statistical software packages). This is precisely a test for overall joint significance of the regression. We estimate the following model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

We formulate the following joint null hypothesis (H_0) stating that the regressors have jointly no effect on the dependent variable:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

The alternative hypothesis (H_1) is:

$$H_1 : H_0 \text{is not true}$$

The formula for the F statistic (or F ratio), where q is the number of restrictions (in this example we are imposing three restrictions), and $n-k-1$ is the number of degrees of freedom of the unrestricted model, is defined by:[1]

$$F = \frac{SSR_r - SSR_{ur}}{SSR_{ur}} * \frac{(n-k-1)}{q}$$

First, we estimate the *unrestricted* model. With this term, we mean the entire or complete model:

```
reg1 <- lm(wage1$wage ~ wage1$educ + wage1$exper + wage1$tenure)
```

We can now calculate the sum of squared residual (SSR) of the unrestricted model:

```
SSR_ur <- sum(reg1$residuals^2)
```

We then estimate the *restricted* model. The restricted model has clearly less parameters than the unrestricted model. Since we are performing an F test for the overall significance of the regression, we must regress the dependent variable *wage* on just an intercept. In R, this is done by including only a “1” after the *tilde* sign in the *lm* function.

[1] The number of degrees of freedom of the unrestricted model is given by $n-k-1$ where n is the number of observations, k is the number of independent variables and 1 stands for the coefficient of the intercept.

```r
reg2 <- lm(wage1$wage ~ 1)
```

We can now calculate the SSR of the restricted model.

```r
SSR_r <- sum(reg2$residuals^2)
```

We report the results in a single table created using the *stargazer* package.

```r
stargazer(reg1, reg2,
          type = "latex",
          header = FALSE,
          title = "F test for the overall significance of the regression.",
          keep.stat = c("n", "rsq", "f"))
```

Table 1: F test for the overall significance of the regression.

	Dependent variable:	
	wage	
	(1)	(2)
educ	0.599***	
	(0.051)	
exper	0.022*	
	(0.012)	
tenure	0.169***	
	(0.022)	
Constant	-2.873^{***}	5.896***
	(0.729)	(0.161)
Observations	526	526
R^2	0.306	0.000
F Statistic	76.873*** (df = 3; 522)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Finally, we can calculate the F statistic and its corresponding p-value. We compare the value of our F statistic (and its p-value) with the value provided by R (see the last row of the first column in the table above).

```r
df_ur <- reg1$df.residual # 522

df_r <- reg2$df.residual # 525

q <- df_r - df_ur # 3

F_test <- (SSR_r - SSR_ur) / SSR_ur * df_ur / q
F_test
```

```
## [1] 76.87317
```

```r
pval <- pf(F_test, q, df_ur, lower.tail = FALSE)
pval
```

```
## [1] 3.405862e-41
```

We choose a significance level (α) of 1% and calculate the corresponding critical value in the F distribution.

```
qf(0.01, df = 3, df2 = 522, lower.tail = FALSE)
```

```
## [1] 3.819327
```

What is the conclusion of the test? We can observe that our F value is clearly larger the critical value for the chosen significance level of 1%. Our p-value is also very very small, certainly smaller than the significance level of 1%. We can therefore soundly reject the null hypothesis that the variables are not jointly significant. We can also create the graph of the F distribution. In green we mark the rejection region for the significance level that we have choosen.