Robust F-test

November 26, 2013

We learned this week about the F-Ratio and Anova. Again, as explained in Field u. a. (2012) this Anova procedere can also be seen through the eyes of regression. We can compare two models, where one is "'nested"' in the other. Nested means that if you have two model U and R with p_U and p_R parameters and R is nested in U, than $p_U > p_R$. For this we look at the squared sum of residuals of the two models:

$$F_{q,n-k-1} = \frac{(RSSR - USSR)/q}{\frac{USSR}{n-k-1}}$$

Where USSR is the sum squared residuals of the <u>u</u>nrestricted model (the "'bigger"' model where the coefficients are estimated) and the RSSR the sum squared residuals of the (smaller) <u>r</u>estricted model and q is the number of restrictions. For example if you have a regression with a dependend variable and 2 independent variables (restricted) and another with the same dependend but 4 independent variables (unrestricted), you can compare this two models with the F statstic above. This tests whether the two additional variables are zero.
Again in Anova the procedure is just the same, we compare two models by decomposing the variance: We have on one hand a model, the unrestricted Modell (U) where we take the group means. This is a model with the same number of parameters as groups. On the other hand we have the restricted model where we just take the grand mean (R) with zero parameters. So if we use the notation of the book, with $SS_T = SS_R + SS_M$ and q as the number of groups, we again have:

$$F_{q,n-k-1} = \frac{(RSSR - USSR)/q}{\frac{USSR}{n-k-1}} = \frac{(SS_T - SS_R)/q}{\frac{SS_R}{n-k-1}} = \frac{MS_M}{MS_R}$$

Because $RSSR \equiv SS_T$, the total variance (since the restricted model is only the mean) and $USSR \equiv SS_R$, the variation within the groups (because the unrestricted model constists of the group means).
Yet this approach has its disadvantages, mainly the assumption of homgeneity of variance. In an Anova analysis this means, that the compared groups should have equal

variances. In the regression context it means that the variance of the error term should be equal for every value of the dependent variable (called Homoskedastie). This is a big problem, because in many cases this assumption does not hold.
So in the basic course for econometrics for economists, we learned that there is a better way to calulculate an F-test from the book of Stock and Watson[1](Stock u. Watson, 2012). This kind of test does not use the variance of the error to compare models. It rather works very similar to a t test. A t test uses the estimate substracts the value it should have under the null hypothesis and divides this difference by the obtained standard error of the estimate. Here again, the key is the sampling distribution.
So if we want to test whether a regression coefficient is likely to be a certain value β_j, we use:

$$t = \frac{\hat{\beta} - \beta_j}{\widehat{SE(\beta)}}$$

The heteroskedastierobust F statistic now looks like this:

$$F_{q,\infty} = (R\hat{\beta} - r)'(R\Sigma_\beta R')^{-1}(R\hat{\beta} - r)/q$$

Let's go through this;
β is a $((k+1) \times 1)$ vector, where k is the number of regressors. So you have $\beta = \begin{pmatrix}\beta_0 & \beta_1 & \dots & \beta_k\end{pmatrix}'$.
R is a $(q \times (k+1))$ Matrix, where k+1 is again the number of β's in the regression. The q is a very important part of the analysis and denotes the number of restrictions. For example if you test wheter three coefficients are jointly zero, than $q = 3$, if you test whether two coefficients are the same, $q = 1$. So the columns of R denote the number of β's and the rows the number of resctrictions. This is important, because to calculate the F stastistic above, the matrices and vectors have to match. r is just a $(q \times 1)$ vector, where q is again the number of restrictions.
Finally $\widehat{\Sigma_\beta}$ is the estimated variance-covariance matrix, which is simply a matrix with all variances and covariances of the coefficients:

$$\widehat{\Sigma_\beta} = \begin{pmatrix} var(\beta_0) & cov(\beta_0, \beta_1) & \dots & cov(\beta_0, \beta_k) \\ cov(\beta_1, \beta_0) & var(\beta_1) & \dots & cov(\beta_1, \beta_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\beta_k, \beta_0) & var(\beta_1, \beta_0) & \dots & var(\beta_k) \end{pmatrix}$$

For someone who has never seen this statistic before, it is best to make an example.
First I will use the example I also made in the R script to this text: It is a dataset, called "'data"' with four variables. I regress one variable on the other three, so that I have a model with three independent variable:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

[1] The R script "'Ftest"' includes a simple function with which you can relatively easy calculate this statistic and obtain a p-value.

I now want to test, wheter either β_0 and β_1 or β_3 and β_4 are in fact the same. This is a test with two restrictions:

$$H_0 : \beta_0 = \beta_1 \ , \ \beta_2 = \beta_3$$

We can rewrite this with R, r and β, so that $H_0 : R\beta = r$:

$$H_0 : \underbrace{\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}}_{R} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ 0 \end{pmatrix}}_{r}$$

So I have determined the model itself, R, r and we know that $q = 2$. I can now insert this data into my function and it will do the rest of the calculations for me:

```
R ←matrix(c(1,0,-1,0,0,1,0,-1), ncol=4,nrow=2)
r ←matrix(c(0,0),ncol=1,nrow=2)
Ftest(lm(score∼age+gender+nerv),R,r,2)
```

You can look at the output by running the R script.
Second it can be shown, that this complex looking F statistic is basically the same as the t-statistic:
For this purpose we make a test with only one restriction, so that $q = 1$. For example we take again a regression with k independent variables:

$$y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i$$

And now make following test:

$$H_0 : \beta_j = 0 \ , \ H_1 : \beta_j \neq 0$$

or in terms of r, R and beta, where R has one row with 1 at R_{1j} and zeros otherwise:

$$H_0 : \underbrace{\begin{pmatrix} 0 & 0 & \dots & 1 & \dots & 0 & 0 \end{pmatrix}}_{R} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \underbrace{0}_{r}$$

So, with q=1:

$$(R\hat{\beta} - r)' = (\begin{pmatrix} 0 & 0 & \dots & 1 & \dots & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} - 0)' = (\beta_j - 0)'$$

$$(R\Sigma_\beta R')^{-1} = [\begin{pmatrix} 0 & 0 & \dots & 1 & \dots & 0 & 0 \end{pmatrix} \begin{pmatrix} var(\beta_0) & cov(\beta_0,\beta_1) & \dots & cov(\beta_0,\beta_k) \\ cov(\beta_1,\beta_0) & var(\beta_1) & \dots & cov(\beta_1,\beta_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\beta_k,\beta_0) & var(\beta_1,\beta_0) & \dots & var(\beta_k) \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}]^{-1}$$

$$= [var(\beta_j)]^{-1}$$

Finally we have:

$$F_{q,\infty} = (\beta_j - 0)'[var(\beta_j)]^{-1}(\beta_j - 0) = \frac{\beta_j^2}{var(\beta_j)} = t^2$$

So the F statistic for $q = 1$ is simple the squared t statistic. It is basically the same principle, only with the joint distribution of the estimated coefficients, which is why we have to use the Σ_β-Matrix. Again the key of those calculations is, that we use the sampling distribution of our coefficients to determine the F statistic, just as we do it for the t statistic. This makes the statistic independent of the variance of the error term. Again if we assume to have a big enough sample, that the central limit theorem holds (I have another text on that on olat, see "'Central Limit Theorem"') we also know a lot of the joint distribution of the coefficients. So if our sample is large and since the assumption of homogeneity of variance is not needed, we can use this F statistic in a wide range of applications. In addition it is a lot easier to make tests with this statistic than with the homoskedastie-only statistic above, where you are just able to compare models.

References

[Field u. a. 2012] Field, Andy ; Miles, Jeremy ; Field, Zoe: *Discovering Statistics using R*. SAGE publications, 2012

[Stock u. Watson 2012] Stock, James H. ; Watson, Mark M.: *Introduction to Econometrics*. Pearson, 2012