Econometrics I TA Session

Wang Xin

June 16, 2022

Contents

1 Simplifing the F Test

As mentioned before, the null hypothesis can be shown as $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$.

Under this restriction, we can rewrite the question as

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$
$$s.t.\ \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

We denote the optimal solution of the above equation as $\tilde{\boldsymbol{\beta}}$,

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{r} - \mathbf{R}\hat{\boldsymbol{\beta}})$$

$$\tilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$$
$$(\mathbf{R}\hat{\boldsymbol{\beta}} - r)'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - r) = \tilde{\mathbf{u}}'\tilde{\mathbf{u}} - \mathbf{e}'\mathbf{e}$$

Moreover, the coefficient of determination of the restricted model and unrestricted model are

$$\tilde{R}^2 = 1 - \frac{\tilde{\mathbf{u}}'\tilde{\mathbf{u}}}{\mathbf{y}'\mathbf{M}\mathbf{y}} \quad \hat{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}\mathbf{y}}$$

Therefore, the F statistics could be simplified,

$$\frac{\tilde{\mathbf{u}}'\tilde{\mathbf{u}} - \mathbf{e}'\mathbf{e}/G}{\mathbf{e}'\mathbf{e}/(n-k)} = \frac{(\hat{R}^2 - \tilde{R}^2)/G}{(1-\hat{R}^2)/(n-k)} \sim F(G, n-k) \tag{1}$$

where $\mathbf{M} = \mathbf{I}_n - \frac{1}{n}\mathbf{i}\mathbf{i}'$ and $\mathbf{i} = (1, 1, \cdots, 1)'$.

2 A Simple Example

We consider the following model that explains major league baseball players' salaries:

$$\begin{aligned}\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg \\ +\beta_4 hrunsyr + \beta_5 rbisyr + u\end{aligned} \tag{2}$$

where *salary* is the 1993 total salary, *years* is years in the league, *gamesyr* is average games played per year, *bavg* is career batting average (for example, *bavg* 5 250), *hrunsyr* is home runs per year, and *rbisyr* is runs batted in per year.

Suppose we want to test the null hypothesis that, once years in the league and games per year have been controlled for, the statistics measuring performance—*bavg*, *hrunsyr*, and *rbisyr*—have no effect on salary. Essentially, the null hypothesis states that productivity as measured by baseball statistics has no effect on salary.

In terms of the parameters of the model, the null hypothesis is stated as

$$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$$

From the data gathered from researchers, the model can be estimated as,

$$\begin{aligned}\widehat{\log(salary)} &= 11.9 + 0.689 years + 0.126 gamesyr + 0.00098 bavg \\ &\quad (0.29) \quad (0.0121) \qquad (0.0026) \qquad\quad (0.00110) \\ &\quad + 0.0144 hrunsyr + 0.0108 rbisyr \\ &\quad\quad (0.0161) \qquad\quad (0.0072) \\ n &= 353,\ SSR = 183.186,\ R^2 = 0.6278,\end{aligned} \tag{3}$$

where SSR is the sum of squared residuals, which is $\mathbf{e}'\mathbf{e}$ in equation (1).

Knowing the sum of squared residuals in (2) tells us nothing about the truth of the null hypothesis. However, the factor that will tell us something is how much the SSR increases when we drop the variables *bavg*, *hrunsyr*, and *rbisyr* from the model. Remember that, because the OLS estimates are chosen to minimize the sum of squared residuals, the SSR always increases when variables are dropped from the model; this is an algebraic fact. The question is whether this increase is large enough, relative to the SSR in the model with all of the variables, to warrant rejecting the null hypothesis.

The model without the three variables in question is simply

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr \tag{4}$$

In the context of hypothesis testing, equation (4) is the restricted model for testing H_0; model (2) is called the unrestricted model. When we estimate the restricted model, we obtain

$$\begin{aligned}\widehat{\log(salary)} &= 11.22 + 0.0713 years + 0.0202 gamesyr \\ &\quad (0.11) \quad (0.0121) \qquad (0.0013) \\ n &= 353,\ SSR = 198.311,\ R^2 = 0.5971,\end{aligned} \tag{5}$$

We now can easily compute $(SSR_r - SSR_{ur})/SSR_{ur}$ and to multiply the result by $(n-k)/G$; the reason the formula is stated as in (1) is that it makes it easier to keep the numerator and denominator degrees of freedom straight. Using the SSRs in (3) and (5), we have

$$F = \frac{(198.311 - 183.186)}{183.186} \cdot \frac{347}{3} \approx 9.55. \tag{6}$$

This number is well above the 1% critical value in the F distribution with 3 and 347 degrees of freedom, and so we soundly reject the hypothesis that *bavg*, *hrunsyr*, and *rbisyr* have no effect on salary.

3 Relationship between F and t Statistics

Come back to equation (3), it reveals that, whereas years and *gamesyr* are statistically significant, none of the variables *bavg*, *hrunsyr*, and *rbisyr* has a statistically significant t statistic against a two-sided alternative, at the 5% significance level. (The t statistic on *rbisyr* is the closest to being significant; its two-sided p-value is 0.134.) Thus, based on the three t statistics, it appears that we cannot reject H_0.

This conclusion turns out to be wrong. The outcome of the joint test may seem surprising in light of the insignificant t statistics for the three variables. What is happening is that the two variables *hrunsyr* and *rbisyr* are highly correlated, and this multicollinearity makes it difficult to uncover the partial effect of each variable; this is reflected in the individual t statistics. The F statistic tests whether these variables (including *bavg*) are jointly significant, and multicollinearity between *hrunsyr* and *rbisyr* is much less relevant for testing this hypothesis.

The F statistic is often useful for testing exclusion of a group of variables when the variables in the group are highly correlated. For example, suppose we want to test whether firm performance affects the salaries of chief executive officers. There are many ways to measure firm performance, and it probably would not be clear ahead of time which measures would be most important. Since measures of firm performance are likely to be highly correlated, hoping to find individually significant measures might be asking too much due to multicollinearity. But an F test can be used to determine whether, as a group, the firm performance variables affect salary.

In this example regressions, two (or more) variables that each have insignificant t statistics can be jointly very significant. It is also possible that, in a group of several explanatory variables, one variable has a significant t statistic but the group of variables is jointly insignificant at the usual significance levels. What should we make of this kind of outcome? For concreteness, suppose that in a model with many explanatory variables we cannot reject the null hypothesis that β_1, β_2, β_3, β_4, and β_5 are all equal to zero at the 5% level, yet the t statistic for $\hat{\beta}_1$ is significant at the 5% level. Logically, we cannot have $\beta_1 \neq 0$ but also have β_1, β_2, β_3, β_4, and β_5 all equal to zero! But as a matter of testing, it is possible that we can group a bunch of insignificant variables with a significant variable and conclude that the entire set of variables is jointly insignificant. (Such possible conflicts between a t test and a joint F test give an example of why we

should not "accept" null hypotheses; we should only fail to reject them.) The F statistic is intended to detect whether a set of coefficients is different from zero, but it is never the best test for determining whether a single coefficient is different from zero. The t test is best suited for testing a single hypothesis. (In statistical terms, an F statistic for joint restrictions including $\beta_1 = 0$ will have less power for detecting $\beta_1 \neq 0$ than the usual t statistic.)

Unfortunately, the fact that we can sometimes hide a statistically significant variable along with some insignificant variables could lead to abuse if regression results are not carefully reported. For example, suppose that, in a study of the determinants of loan-acceptance rates at the city level, x_1 is the fraction of black households in the city. Suppose that the variables x_2, x_3, x_4, and x_5 are the fractions of households headed by different age groups. In explaining loan rates, we would include measures of income, wealth, credit ratings, and so on. Suppose that age of household head has no effect on loan approval rates, once other variables are controlled for. Even if race has a marginally significant effect, it is possible that the race and age variables could be jointly insignificant. Someone wanting to conclude that race is not a factor could simply report something like "Race and age variables were added to the equation, but they were jointly insignificant at the 5% level."