[STAT-11] More on testing linear models

Miguel-Angel Canela
Associate Professor, IESE Business School

Nested models

You will find in the literature that, frequently, the objective of the statistical analysis is to choose between two **nested models**. We say that two models are nested when one is a particular case of the other, resulting from constraining some of the parameters of the general model. This concept is easily understood in an example of multiple linear regression. Imagine the linear regression models associated to the equations

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \qquad Y = \beta_0 + \beta_1 X_1 + \epsilon.$$

It is clear that the **restricted model** (right) is obtained from the **unrestricted model** (left) by constraining $\beta_2 = \beta_3 = 0$. When we talk loosely about comparing these two models, we usually mean testing, in the unrestricted model, the null $H_0 : \beta_2 = \beta_3 = 0$. The equalities that form this null hypothesis are called **restrictions** or constraints. I present in this lecture an F test for linear restrictions in the context of OLS estimation. There is a similar approach for maximum likelihood estimation, the **likelihood ratio test**, which you will use in the Econometrics and Multivariate Stats courses.

Testing linear restrictions

Any set of linear restrictions can be tested in a simple way. The idea of the test is as follows. The OLS estimates are the solution of an optimization problem. In the restricted model, the search for the minimum is carried out in a smaller set. For instance, in the example of the preceding section, setting $\beta_2 = \beta_3 = 0$ restricts the search to a two-dimensional space, instead of the four-dimensional space of the unrestricted model. This implies that the residual sum of squares SSR increases when introducing the restrictions and, therefore, R-squared decreases. The test statistic can be written in terms of either SSR or R-squared.

Let the subscripts r and u refer to the restricted and the unrestricted model, respectively. Then, the test statistic is

$$F = \frac{\dfrac{\mathrm{SSR_r} - \mathrm{SSR_u}}{\mathrm{df_r} - \mathrm{df_u}}}{\dfrac{\mathrm{SSR_u}}{\mathrm{df_u}}} = \frac{\dfrac{R^2_\mathrm{u} - R^2_\mathrm{r}}{\mathrm{df_r} - \mathrm{df_u}}}{\dfrac{1 - R^2_\mathrm{u}}{\mathrm{df_u}}}\,.$$

Under the null (the two restrictions), this statistic has an $F(\mathrm{df_r} - \mathrm{df_u}, \mathrm{df_u})$ distribution, which is used to calculate the corresponding P-value. Note that, in the above example, the unrestricted model has $\mathrm{df_u} = n - 4$, and the restricted model $\mathrm{df_r} = n - 2$, so the distribution of the F statistic is $F(2, n - 4)$.

We typically regard this F test as a test on the increase in R-squared, typically reported as ΔR^2. In certain fields, it is customary to report two or more nested models, testing the increment ΔR^2 due to the additional variables. This is sometimes called **hierarchical regression**, not to be confounded with the hierarchical linear models which are another name of the multilevel models that will be discussed in the Econometrics course.

Particular cases

This F test gives as particular cases the tests of the preceding lecture. When testing only one term, it is equivalent to the corresponding t test that comes in the coefficients table. When testing all the terms except the intercept, it is the F test associated to the ANOVA table. More specifically, in the 4-terms unrestricted model used above as an example:

- The default F test applies to the null $\beta_1 = \beta_2 = \beta_3 = 0$.
- The t test for the null $H_0 : \beta_j = 0$ ($j = 1, 2, 3$) is equivalent to the corresponding F test for this restriction. The square of the t statistic equals the F statistic.

Example 1. Models for the influence of education on wages are frequently used in Econometrics courses. We have already seen one of these examples. The `wage1` data set is a subset of a bigger data set used in the Wooldridge's textbook. It contains 526 observations on wages of workers. The variables included in the discussion of this example are:

- `wage`, 1976 wages in US work force, in US dollars per hour.
- `educ`, years of schooling. `educ = 12` corresponds to complete high school education.
- `exper`, years of potential experience.
- `tenure`, years with current employer.

A standard approach would be to run a regression of wages, in log scale on the three explanatory variables. Table 1 is the table of coefficients. We have $R^2 = 0.316$ ($F = 80.4$, $P < 0.001$). The coefficient of `educ` is significant, so the conclusion is easy.

TABLE 1. Linear regression results (Example 1)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	0.284	0.104	2.79	0.007
educ	0.092	0.007	12.6	0.000
exper	0.004	0.002	2.39	0.017
tenure	0.0221	0.003	7.13	0.000

I take first the 3-regressor model as the unrestricted model and the model without `exper`, whose coefficient looks less significant than the other two, as the restricted model. Table 2 is the new table of coefficients.

TABLE 2. First restricted model

Coefficient	Estimate	Std. error	t value	p-value
Intercept	0.404	0.092	4.41	0.000
educ	0.087	0.007	12.4	0.000
tenure	0.026	0.003	9.63	0.000

R-squared falls to 0.308 ($\Delta R^2 = 0.008$). The associated F statistic (df $= (1, 522)$) is

$$F = \frac{0.316 - 0.308}{(1 - 0.316)/522} = 5.72,$$

which is the square of the t statistic associated to the coefficient of `exper` in Table 1 ($t = 2.39$). The p-values associated are the same.

In a second comparison, I take as the restricted model the **null model** obtained by dropping all the regressors, with $R^2 = 0$. The equation includes only the intercept, which equals the mean

log wages (Table 3). Now, $F = 80.39$ ($P < 0.001$), which is, precisely, the F statistic of the unrestricted model.

TABLE 3. Second restricted model

Coefficient	Estimate	Std. error	t value	p-value
Intercept	1.623	0.023	70.0	0.000

In a final exercise, I test two coefficients (this test is not included in the default regression output). I take as the restricted model the equation obtained by dropping `exper` and `tenure` (Table 4). Now, $R^2 = 0.186$ ($\Delta R^2 = 0.130$, $F = 49.7$, $P < 0.001$).

TABLE 4. Third restricted model

Coefficient	Estimate	Std. error	t value	p-value
Intercept	0.584	0.097	5.60	0.000
educ	0.083	0.008	10.9	0.000

¶ Source: JM Wooldridge (2013), *Introductory Econometrics — A Modern Approach*, South-Western College Publishing.

Multicollinearity

Let me come back to the notation of the preceding lectures and consider an equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$. There is a relatively simple way to write thew variance of an estimator $\hat{\beta}_j$. Denoting by R_j^2 the R-squared statistic associated to the regression of X_j on the other independent variables and by s_j^2 the (sample) variance of X_j, the formula is

$$\operatorname{var}[\hat{\beta}_j] = \frac{\sigma^2}{(n-1)s_j^2(1-R_j^2)} \,.$$

The squared root of this variance is the standard error $\operatorname{se}[\hat{\beta}_j]$, which is used for testing purposes. This formula shows that, if we increase the sample size, holding more or less constant the dispersion of X_j, the standard error tends to zero. So, $\hat{\beta}_j$ is a consistent estimator. But it also shows how the presence of the other independent variables affects the standard error. When R_j^2 is close to 1, we say that X_j is affected by multicollinearity. In a multicollinearity situation, the standard error could be high (so, the t statistic could be low, losing significance).

A popular measure of multicollinearity is the **variance inflation factor** (VIF), defined as

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2} \,.$$

The VIF tells us how much larger the variance of $\hat{\beta}_j$ is, compared to what it would be if X_j were uncorrelated with the other independent variables. The standard error is increased by a factor equal to the square root of the VIF. There is no consensus on the threshold for the VIF, although 10 is a popular choice.

There is a certain confusion about multicollinearity. Many people believe that it stops the regression being "correct". This is not so. It is equally correct, although it may fail to prove the existence of some effect. Also, some "detect" potential multicollinearity just because a pair of variables are

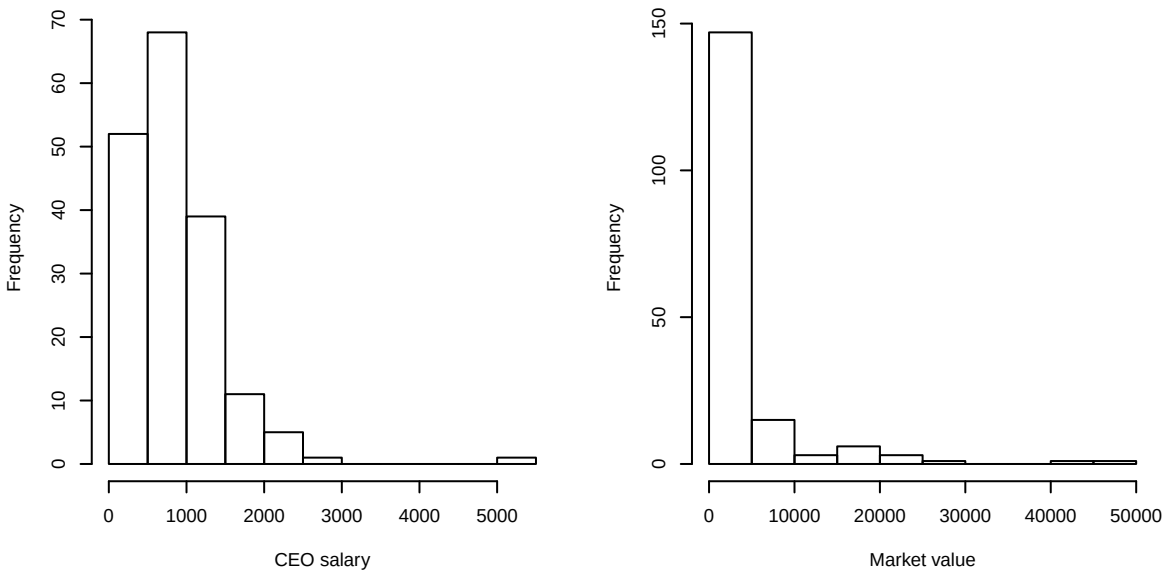

Figure 1. Distribution of raw data (Example 2)

strongly correlated. To see why this is not so, consider a regression with two independent variables whose correlation is 0.9. The above formula would give VIF = 5.26, not that strong.

But there is a context in which you should foresee a multicollinearity issue. Imagine that you include, besides a variable X, a squared term X^2. When the range of X is far from zero, X and X^2 can be (very) strongly correlated. If we test the effect of X, the multicollinearity can sweep away the significance. As an illustration, take a variable called *year*, with values from 1930 to 1960. Then, $\text{cor}[year, year^2] = 0.999998$, so VIF = 25.0. Even subtracting the time origin, the problem persists: $\text{cor}[year, (year - 1930)^2] = 0.966335$ (VIF = 15.1). Centering is always the best policy, since $\text{cor}[year, (year - 1945)^2] = 0$. A similar problem may occur with the product terms that we use to test moderation effects (see next lecture).

Example 2. The relationship between firm performance and CEO salary has been the subject of much discussion, not only in the academic context, but also in the media. The `ceosal2` data set is a subset of one used in Wooldridge's Econometrics textbook, based on cross-sectional firm-level data. The variables are:

- `salary`, CEO 1990 compensation, in thousands of US dollars.
- `sales`, 1990 firm sales, in millions of US dollars.
- `mktval`, market value at the end of 1990, in millions of US dollars.
- `profits`, 1990 profits, in millions of US dollars.

I use log scale for all variables except `profits` (it takes negative values). The log scale can be justified by the skewness of the distributions, or by the presence of extreme values (see Figure 1). Moreover, in econometric analysis, log transformations are typically applied to salary and company size.

TABLE 5. Correlation matrix (Example 2)

	log(sales)	log(mktval)	profits
log(salary)	0.530	0.481	0.397
log(sales)		0.736	0.606
log(mktval)			0.777

Table 5 is the correlation matrix. As expected, all the correlations are positive. The potential drivers of salary are positively related to it and, moreover, they are strongly correlated among them, raising a concern about multicollinearity.

I start the regression analysis with a regression line. Table 6 is the coefficients table. The slope of this line provides an assessment of the effect of market value on the CEO's salary. But it is not clear whether this is due to market value alone, or to other financial aspects that are positively correlated to market value. For instance, since one expects company sales and market value to be positively correlated, we may wonder which would be effect of a change in the market value, holding sales constant.

TABLE 6. Regression line ($R^2 = 0.232$)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	4.678	0.265	17.6	0.000
log(mktval)	0.257	0.035	7.27	0.000

My first multiple regression model (Table 7) includes `log(sales)` as a control variable. Both coefficients are positive, as expected, and significant. The coefficient of `log(mktval)` is lower than in Table 6, as expected. Which of them is the right one? From the statistical point of view, both are. Nevertheless, for an econometrician, who thinks in causal terms, the second equation could be right, but the first one is wrong.

TABLE 7. Regression with control variable ($R^2 = 0.299$)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	4.621	0.254	18.2	0.000
log(sales)	0.162	0.039	4.09	0.000
log(mktval)	0.107	0.050	2.13	0.035

A second multiple regression model (Table 8) includes `profits` as a second control variable. We do not find a relevent change in R-squared, but `profits` steals part of the effect of `log(mktval)`. There is no contradiction here, since the coefficient of `log(mktval)` has now a different interpretation.

TABLE 8. Regression with two control variables ($R^2 = 0.299$)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	4.687	0.380	12.3	0.000
log(sales)	0.161	0.040	4.04	0.000
log(mktval)	0.098	0.064	1.53	0.128
profits	3.57e-05	1.52e-04	0.23	0.815

The inclusion of `profits` in the equation is questionable, since its contribution to the R-squared is irrelevant. Also, the strong correlations suggests a potential multicollinearity problem. But I leave this for the homework.

¶ Source: JM Wooldridge (2013), *Introductory Econometrics — A Modern Approach*, South-Western College Publishing.

Homework

A. In Example 1, that experience has a positive effect on wages seems evident, but that the effect is linear is unclear. Common sense tells us that the first years are more relevant, and the effect gets weaker when workers have experience enough. A simple way to account with the curvilinear effect of the experience on the wages is to include a squared term in the equation. Try this, testing the two terms related to the experience in one shot.

B. Calculate the VIF for the three independent variables of Table 8. Is there a multicollinearity issue?