[STAT-20] Testing regression coefficients

Miguel-Angel Canela
Associate Professor, IESE Business School

The normality assumption

This lecture deals with testing the coefficients in a linear regression equation

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon.$$

For testing, distributional assumptions (not used in the preceding lecture) are needed. The standard assumption is that the error term is normally distributed. We write this as $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This looks quite straightforward, but a bit of confusion is typical around this point, so I include here some remarks:

- It is ϵ, not Y, what is assumed to be normal. This means that, for a fixed X point, Y has a normal distribution. For another point, Y will also be normal, but with a different mean. This is sometimes expressed saying that Y is conditionally normal. An example may help to clarify. Take as Y the income in a certain population and as X a dummy for gender,and suppose that the average income is different in the two subpopulations. The distribution of Y, conditional to $X = 1$, is assumed to be normal, and the same for the distribution of Y conditional to $X = 0$, which means that the income is normally distributed in the two subpopulations. But when the two subpopulations are merged, the distribution is no longer normal. We will have a "camel" distribution here.
- Testing a coefficient usually means testing a null $H_0: \ \beta_i = 0$. When the null is rejected, we say that that coefficient is significant. Although a p-value is typically reported by stat packages for β_0, we practically never test the intercept, which we leave in the equation, irrespective of its significance.
- Many people believe that a nonsignificant coefficient means that the corresponding term *should* be dropped. It is not so. A better rule would be that it *could* be dropped.

What can be done without the normality assumption? Practically the same, for big samples. Indeed, it can be shown that the OLS estimators of the regression coefficients are asymptotically normally distributed, so that most of the previous discussion remains valid.

The t tests

Under the normality assumption, we can calculate $1 - 2\alpha$ confidence limits for the regression coefficients with the formula $\hat\beta \pm t_\alpha \,\hat{\text{se}}[\hat\beta]$. Standard errors can also be used to run a t test. For the null $H_0: \ \beta = 0$, we use

$$t = \frac{\hat\beta}{\hat{\text{se}}[\hat\beta]}\,,$$

with $\text{df} = n - k - 1$. Note that the standard error is not the exact value, but an estimate. That is why we use the Student t instead of the standard normal. In the computer, the standard regression output contains, besides every parameter estimate, the standard error, the t statistic and the p-value (Table 1). Also, note that, for very big samples, t statistics are likely to be significant.

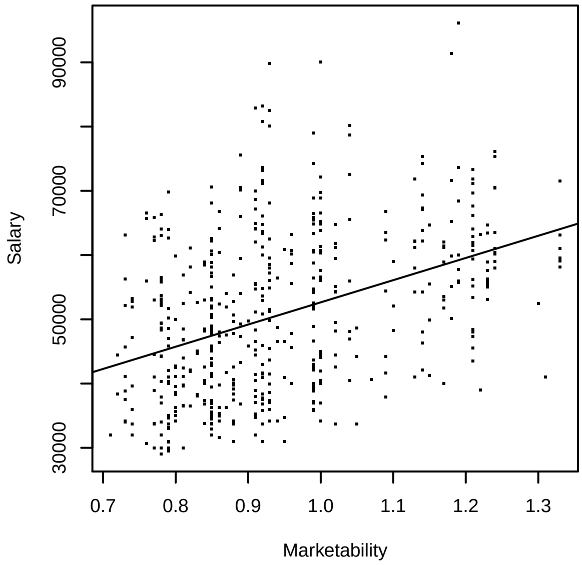

Figure 1. Regression line (Example 1)

The F test

In many stat packages, the default regression report includes the ANOVA decomposition (it is so in Stata, but not in `R`). The F statistic associated to this ANOVA decomposition,

$$F = \big(n-2\big)\,\frac{\text{SSE}}{\text{SSR}} = \frac{(n-2)\,R^2}{1-R^2}$$

is used to test the null $H_0: \ \beta_1 = \cdots = \beta_k = 0$.

The second expression of this statistic, involving R^2, shows that significance occurs when R^2 is close enough to 1. We say then that R^2 is significant, or that the (multiple) correlation is significant. It is also obvious, from the formula, that weak correlations are significant with samples big enough. In the simple regression case, the F statistic is the square of the t statistic associated to β_1, so it is redundant (this is no longer true in multiple regression).

Residual analysis

The analysis of the residuals is useful for checking the validity of the model. This analysis is similar to that of the residuals of the one-way ANOVA (they are a particular case of regression residuals, see lecture 21), plus the possibility of a **residual plot**. In a residual plot, we place the residuals (standardized or not) in the ordinates, and x_i, the predicted values $\hat{y}_i$ or the order in which the data were obtained, in the abscissas.

Example 1. The `market` data set contains data from a study on the salaries of academic staff in Bowling Green State University. This data set has been used in several textbooks and can be considered as a standard example. The sample size is 514. We use here the variables:

- `salary`, academic year (9 month) salary in US dollars.
- `market`, marketability of the discipline, defined as the ratio of national average salary paid in the discipline to the national average across disciplines.

It is natural here to take marketability as the independent variable, fitting a regression line to the data to produce an equation for predicting the salary from the marketability. The R-squared

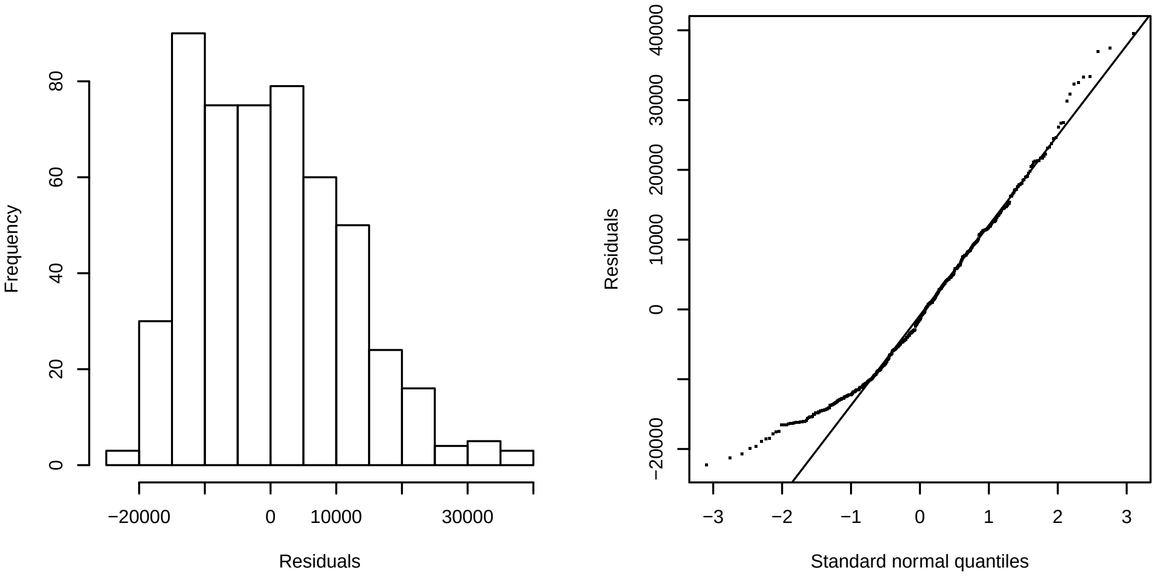

Figure 2. Distribution of residuals (Example 1)

statistic is $R^2 = 0.166$. In the coefficients table (Table 1), we find the coefficient estimates, the standard errors, the t statistic and the p-value.

TABLE 1. Linear regression results (Example 1)

Salary	Coefficient	Std. error	t statistic	P-value
Marketability	34,545.2	3,424.3	10.09	0.000
Constant	18,097.0	3,288.0	5.50	0.000

The F statistic is $F(1, 512) = 101.8$. Note that this is the same as the square of the t statistic for marketability ($t = 10.09$), with the same P-value.

The diagnostic plots of Figure 2 reveal a clear departure from normality on the left tail. Indeed, the skewness is Sk = 0.595, the kurtosis K = −0.017, and the Jarque-Bera statistic JB = 30.3 ($P < 0.001$). Nevertheless, given the sample size, non-normality is not a problem here.

Homework

A. Rerun the analysis of Example 1 replacing `salary` by `log(salary)`. How do you interpret the coefficients in both cases?