[STAT-22] Regression and correlation (2)

Miguel-Angel Canela
Associate Professor, IESE Business School

Foreword

In a second round on the connection between regression and correlation, this lecture deals with three interesting points:

- We already know that a regression coefficient changes when we include an additional variable in the equation. The change can be more or less relevant, depending on the case. On what does this change depend? As we will see below, we can split an independent variable X in two uncorrelated components, one related to the other independent variables, and one unrelated to them. What matters is how the covariance between Y and X is split between these two components.
- In a simple linear regression analysis, we test the correlation between X and Y by testing the slope coefficient. What about multiple regression? Can we relate the coefficient of X to some kind of correlation between Y and X? Indeed, we can, to a **partial correlation**, partial meaning here that the correlation of Y with the component of X related to the other independent variables has been discounted.
- What happens when we pile up on the right side of a regression equation variables which are strongly correlated? Intuitively, we see this a redundancy, because a variable which is very close to a linear combination of the other independent variables does not contribute to the predictive or explanatory power of the model. So, this variable is less likely to be significant. We call this is a **multicollinearity** issue.

Partitioned regression

Let me adequate a bit the notation, only for this lecture, to simplify the discussion. Y is the dependent variable, X is an independent variable and $Z_1, \dots, Z_k$ are the other independent variables, which we call **control variables**. The coefficient of X can be seen as the effect of X on Y after removing the effect of the control variables. The expression *controlling for* is typically used in this context.

More specifically, I take the complete equation, which I write as

$$Y = b_0 + aX + b_1 Z_1 + \cdots + b_k Z_k + e,$$

and the equation for the regression of X on the controls, which I write as

$$X = c_0 + c_1 Z_1 + \cdots + c_k Z_k + u.$$

This equation can be regarded as the decomposition of X in two components, one related to the controls and another unrelated, the residual term u. So, we take u as the part of X where the influence of the controls has been removed. It can be shown, with a bit of algebra, that the slope of the regression of Y on u is, precisely, the coefficient a (see Example 1).

Partial correlation

The test on the coefficient a can be seen as a test on partial correlation. Suppose that v is the residual term in the regression of Y on $Z_1, \ldots, Z_k$. Then we refer to the correlation of the residual terms u and v as the partial correlation between X and Y, controlling for $Z_1, \ldots, Z_k$.

The partial correlation is taken as a measure of the (linear) association between X and Y after removing the influence of the control variables. So, testing a is the same as testing the partial correlation. Therefore, in a regression equation, we take the significance of a coefficient, not as evidence of correlation between the corresponding independent variable and the dependent variable (as in simple regression), but as evidence of partial correlation. Of course, the partial correlation depends on the set of control variables.

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

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

The squared root of this variance is the standard error $\mathrm{se}[\hat{\beta}_j]$, 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 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, $\mathrm{cor}[\mathit{year}, \mathit{year}^2] = 0.999998$, so VIF = 25.0. Even subtracting the time origin, the problem persists: $\mathrm{cor}[\mathit{year}, (\mathit{year} - 1930)^2] = 0.966335$ (VIF = 15.1). Centering is always the best policy, since $\mathrm{cor}[\mathit{year}, (\mathit{year} - 1945)^2] = 0$. A similar problem may occur with the product terms that we use to test moderation effects (lecture STAT-24).

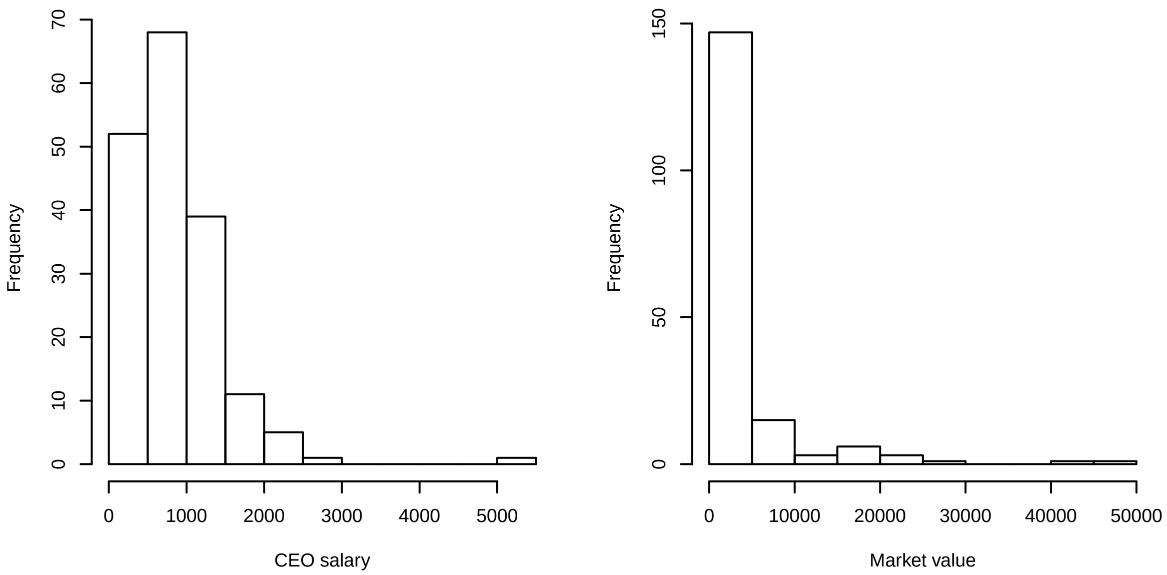

Figure 1. Distribution of raw data (Example 1)

Example 1. 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 1. Correlation matrix

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

Table 1 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 2 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 2. 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 3) 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 2, 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 3. 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

As discussed above, when we test the `log(mktval)` coefficient in Table 2, we are testing the correlation of `log(salary)` and `log(mktval)`. When test the coefficient in Table 3, we are testing their partial correlation, partialling out `log(sales)`. This partial correlation is 0.159, still positive, but much lower than the plain correlation 0.481.

It can be seen in the next regression analysis (Table 4) that the coefficient of `log(mktval)` in the regression of `log(salary)` on `log(lsales)` and `log(mktval)` is the same as the slope in the regression of `log(salary)` on the residuals of the regression of `log(mktval)` on `log(lsales)` (denoted by `resmktval`). This illustrates the interpretation of a regression coefficient as the effect of one variable after removing the effect of the other regressors. The R-squared statistic in Table 4 is the square of the partial correlation.

TABLE 4. Regression after partialling out the control variable $(R^2 = 0.018)$

Coefficient	Estimate	Std. error	t value	p-value
Intercept	6.583	0.045	145	0.000
resmktval	0.107	0.059	1.80	0.073

A second multiple regression model (Table 5) 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 5. 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. Calculate the VIF for the three independent variables of Table 5. Is there a multicollinearity issue?