[STAT-02] Regression and correlation (1)

Miguel-Angel Canela
Associate Professor, IESE Business School

The regression line

I start this lecture refreshing the basics of the regression line. Let us consider a set of n joint observations on two variables X and Y, which we put as the columns $\mathbf{x}$ and $\mathbf{y}$ of a data matrix, and the coefficients b_0 and b_1 of the linear regression equation (of Y on X).

The expression $y = b_0 + b_1 x$ can be taken as the equation of a line, which we call the **regression line**. b_1 is the **slope** and b_0 the **intercept** or constant. In this case, the (matrix) OLS formulas are reduced to simple expressions:

$$b_1 = \frac{\sum \big(x_i - \bar{x}\big)\big(y_i - \bar{y}\big)}{\sum \big(x_i - \bar{x}\big)^2}\,, \qquad b_0 = \bar{y} - b_1 \bar{x}.$$

It follows from the formula of the intercept that $\bar{y} = b_0 + b_1 \bar{x}$, meaning that the regression line crosses the average point $\big(\bar{x}, \bar{y}\big)$. This is equivalent to the sum (and the mean) of the residuals being equal to zero. A consequence of this is that we can write the equation of the regression line as $y - \bar{y} = b_1\big(x - \bar{x}\big)$. This intercept-free version of the regression equation is practical for some analyses.

Regression and correlation

The formulas related to the regression line become more compact if we introduce standard deviations and correlations, which make sense if we take the set of points as a bivariate sample. Dividing by $n - 1$ the numerator and denominator in the expression of the slope, we get

$$b_1 = \frac{s_{xy}}{s_x^2} = r\,\frac{s_y}{s_x}.$$

This tells us that the sample correlation is a standardized regression slope. If $\mathbf{x}$ and $\mathbf{y}$ have unit variance, the slope is equal to the correlation. Now, we can write the regression equation as

$$\frac{y - \bar{y}}{s_y} = r\,\frac{x - \bar{x}}{s_x}\,,$$

which shows that, if both variables are standardized, the slope coincides with the correlation and the intercept is zero. The former is no longer true in multiple regression, where **standardized regression coefficients** are not correlations, although, in most cases, they look as if they were.

Example 1. The `bramex` data set contains the daily returns of the Brazil and Mexico MSCI indexes. It has been extracted from the Datastream database and covers the whole year 2003, with a total of 261 observations (no data in week-ends).

The daily returns are derived from the index values as follows. If x_t is the value of a particular index at day t, the daily return at this day is given by $r_t = x_t/x_{t-1} - 1$. The returns used here come in percentage scale.

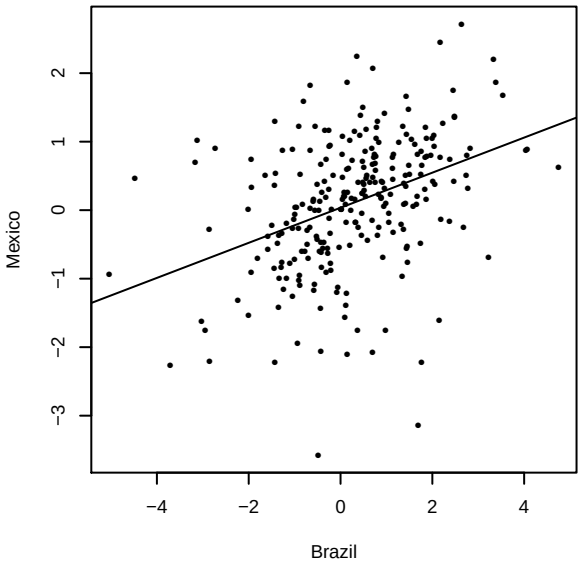

Figure 1. Regression line (Example 1)

Let me denote by $\mathbf{x}$ the vector of Brazil returns and by $\mathbf{y}$ that of Mexico returns. The means are $\bar{x} = 0.273$ and $\bar{y} = 0.105$. The covariance and correlation matrices are, respectively,

$$\mathbf{S} = \begin{bmatrix} 2.120 & 0.543 \\ 0.543 & 0.934 \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} 1 & 0.386 \\ 0.386 & 1 \end{bmatrix}.$$

To calculate the regression line of Mexico on Brazil, I use the formulas of the regression coefficients:

$$b_1 = 0.386 \sqrt{\frac{2.120}{0.934}} = 0.256, \qquad b_0 = 0.105 - 0.256 \times 0.273 = 0.036.$$

I have thus obtained the equation of the line for the regression of Mexico on Brazil,

$$\texttt{mex} = 0.036 + 0.256\,\texttt{bra}.$$

We can see in Figure 1 a scatterplot of these data, with the regression line superimposed.

¶ Source: MA Canela & E Pedreira (2012), Modelling dependence in Latin American markets using copula functions, *Journal of Emerging Markets Finance* **11**, 231–270.

The R-squared statistic

I turn now to general linear regression, where the correlation issue is a bit more complex. I start from the equation

$$\mathbf{y} = b_0\mathbf{1} + b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k + \mathbf{e},$$

which is, in fact a decomposition of $\mathbf{y}$ into two orthogonal vectors. Statisticians prefer to write this as

$$\mathbf{y} - \bar{y} = b_1\big(\mathbf{x}_1 - \bar{x}_1\big) + \cdots + b_k\big(\mathbf{x}_k - \bar{x}_k\big) + \mathbf{e},$$

which is also an orthogonal decomposition. Then, Pythagoras theorem gives us

$$\big\|\mathbf{y} - \bar{y}\big\|^2 = \big\|b_1\big(\mathbf{x}_1 - \bar{x}_1\big) + \cdots + b_k\big(\mathbf{x}_k - \bar{x}_k\big)\big\|^2 + \big\|\mathbf{e}\big\|^2.$$

The left side of the equation is

$$\left\|\mathbf{y} - \bar{y}\right\|^2 = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2.$$

Statisticians call this a **sum of squares**. More specifically, it would be the **total sum of squares**, SST. The first term on the right side is the **sum of squares explained by the regression**, SSE. The last term is the residual sum of squares,

$$\mathrm{SSR} = \sum e_i^2.$$

So, the orthogonal decomposition becomes a formula involving sums of squares, SST = SSE+SSR. This type of formula is called, in Statistics, an **ANOVA decomposition** (this will be later explained). The **residual sum of squares** SSR is a measure of the **goodness-of-fit**. But, in practice, it is difficult to interpret, since it depends on the units of Y and the number of points. So, we use the **R-squared statistic**, defined as follows.

$$R^2 = \frac{\mathrm{SSE}}{\mathrm{SST}}\,.$$

R^2 is taken as the percentage of variation explained by the regression. It is not hard to see that R^2 is the square of the correlation between $\mathbf{y}$ and the vector of **predicted values** $b_0\mathbf{1}+b_1\mathbf{x}_1+\cdots+b_k\mathbf{x}_k$, which is called the **multiple correlation**. An R-squared value close to 1 is taken as an indication of good fit. Nevertheless, how good is the fit that we can expect for a particular type of data is something that we learn only with practice, so it is better to skip general specifications of threshold values for R-squared statistic.

Homework

A. Although the idea of the financial performance of a firm may seem obvious, there is no consensus on how to measure it. Two well known measures are the **return on equity** (ROE) and the **return on assets** (ROA). The ROE measures a firm's efficiency at generating profits from every unit of shareholders' equity. The ROA tells us how profitable the firm's assets are in generating revenue, or, more specifically, how many dollars of earnings it derives from each dollar of assets it controls. The `roeroa` data set covers a wide range of industries. It contains the ROE and the ROA of 426 firms for the year 2000, derived from public sources. The ROE has been calculated as net income over total equity, and the ROA as operating income over total assets. Perform a regression analysis. Which is your conclusion?

B. The `indianbanks` data set contains data on daily opening prices of five Indian banks in the National Stock Exchange (NSE), from 2002-08-12 to 2013-12-31, extracted from Yahoo Finance India. Calculate the correlation matrix of the daily returns and discuss.