

Visualizing the correlation coefficient

*Il coefficiente di correlazione: una sua rappresentazione grafica*¹

Giovanni C. Porzio

Dipartimento Economia e Territorio
Università degli Studi di Cassino
porzio@eco.unicas.it

Luca Scrucca

Dipartimento di Scienze Statistiche
Università degli Studi di Perugia
luca@stat.unipg.it

Riassunto: Il coefficiente di correlazione è uno degli strumenti più utilizzati nell'ambito della statistica per riassumere la relazione tra due variabili, e in particolare il grado della loro interdipendenza. In questo lavoro si propone una metodologia grafica che permette una migliore interpretazione della relazione esistente tra due variabili osservate. Il metodo si basa sull'uso congiunto del diagramma di dispersione, arricchito da una semplice rappresentazione del coefficiente di correlazione, e di un nuovo diagramma costruito a partire dal primo.

Keywords: Linear regression, scatterplot, dynamic graphics.

1. Introduction

The correlation coefficient is probably the most used and taught tool available in statistics to summarize the relationship (and in particular the interdependence) among two random variables X and Y . On the other hand, given a sample of n bivariate observations (x_i, y_i) , the relationship between the two observed variables can be described graphically through a scatterplot of (x_i, y_i) .

In this paper, we aim to enrich information gathered by a scatterplot of a bivariate sample by means of a simple visualization of the sample correlation coefficient. The idea follows common practice in applied statistics, where graphical and numerical analysis are companion. Such an issue has been indeed addressed in the latest years through graphical representation of numerical summaries to be plotted jointly with observed data. As an example, Bowman and Young (1996) proposed to visualize many kinds of hypothesis tests through reference bands to be considered as corresponding acceptance regions. Analogously, we propose to represent the observed data (through a scatterplot and a new plot) along with a visualization of a numerical summary (the sample correlation coefficient).

The method we propose helps the user to evaluate the strength of the correlation among two observed variables. It exploits a simple visualization of the correlation coefficient and a new plot to be built from the two regression lines available in a scatterplot. In a dynamic graphics framework, the method allows to perform a deletion-based influence analysis on the correlation coefficient as well.

¹This work has been supported through M.I.U.R. and C.N.R. funds.

2. The correlation coefficient

The correlation coefficient among two random variables X and Y is defined as

$$\rho = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y},$$

with $\sigma_{X,Y}$ the covariance between X and Y , σ_X and σ_Y the squared root of the variances.

For our purposes, we recall that the correlation coefficient is strictly related to the regression curves (the conditional expected values) of Y against X and conversely of X against Y . Let X and Y be related as

$$E(Y|X = x) = \alpha_{Y|X} + \beta_{Y|X} x \quad \text{and} \quad E(X|Y = y) = \alpha_{X|Y} + \beta_{X|Y} y,$$

and consider drawing the two regression lines $E(Y|X = x)$ and $E(X|Y = y)$ on the same plane, say the y - x plane. Both of them will pass through the marginal expected values $E(Y)$ and $E(X)$, and will have coefficients respectively equal to $\beta_{Y|X}$ and to $1/\beta_{X|Y}$. The angle θ between these two regression lines will be equal to (Kendall and Stuart, 1967)

$$\theta = \arctan \left\{ \frac{1 - \rho^2}{\rho} \frac{\sigma_X \sigma_Y}{\sigma_X^2 + \sigma_Y^2} \right\}.$$

Hence, as ρ varies over its range from -1 to +1, the magnitude of the angle $|\theta|$ increases from 0 to its maximum of $\pi/2$ when $\rho = 0$, and then decreases to 0 again. Thus, the two regression lines coincide if and only if $|\rho| = 1$, while the two lines are at right angles to each other if and only if $\rho = 0$. View from the other side, we have that the correlation ρ is a nonlinear function of the angle θ between the two regression lines.

Finally, we recall that the usual estimator of ρ is the sample correlation coefficient

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{[\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2]^{1/2}} = \frac{s_{xy}}{s_x s_y}$$

with $\bar{x} = \Sigma x_i / n$ and $\bar{y} = \Sigma y_i / n$.

3. A new plot to visualize the coefficient

To gain information on the linear relationship among the observed variables, we propose a new plot in order to visualize the corresponding correlation coefficient.

To draw the plot, consider the two regression lines $E(Y|X)$ and $E(X|Y)$. If estimated via ordinary least squares, they will have equations on the y - x plane:

$$\hat{y} = (\bar{y} - (s_{xy}/s_x^2)\bar{x}) + (s_{xy}/s_x^2)x, \quad \text{and} \quad \tilde{y} = (\bar{y} - s_y^2/s_{xy})\bar{x} + (s_y^2/s_{xy})x. \quad (1)$$

Consider then the vertical distances among these regression lines (1) and the observed values y_i . That is, set $\hat{y}_i = (\bar{y} - (s_{xy}/s_x^2)\bar{x}) + (s_{xy}/s_x^2)x_i$ and $\tilde{y}_i = (\bar{y} - (s_y^2/s_{xy})\bar{x}) + (s_y^2/s_{xy})x_i$, and evaluate the distances $r_i(Y|X) = y_i - \hat{y}_i$ and $r_i(X|Y) = y_i - \tilde{y}_i$.

Let us consider the relationship among these quantities. We show that the OLS regression of $r(Y|X)$ on $r(X|Y)$ has intercept zero and slope equal to the squared correlation coefficient, whereas the OLS regression of $r(X|Y)$ on $r(Y|X)$ has intercept zero and slope equal to one.

Proof. Both intercepts are zero because of the zero mean of the OLS residuals. The slope of the OLS regression of $r(Y|X)$ on $r(X|Y)$ is equal by definition to

$$\frac{\text{Cov}(r(Y|X), r(X|Y))}{\text{Var}(r(X|Y))} = \frac{\text{Var}(r(Y|X))}{\text{Var}(r(X|Y))} = \frac{s_y^2 - s_{xy}^2/s_x^2}{s_y^2(1/r^2 - 1)} = \frac{1 - r^2}{1/r^2 - 1} = r^2.$$

By the same argument, the slope of the OLS regression of $r(X|Y)$ on $r(Y|X)$ is equal to

$$\frac{\text{Cov}(r(Y|X), r(X|Y))}{\text{Var}(r(Y|X))} = \frac{\text{Var}(r(X|Y))}{\text{Var}(r(Y|X))} = 1$$

□

In our study we focus on the regression of $r(Y|X)$ on $r(X|Y)$. This has a convenient form for graphical exploration because the slope of the OLS regression line conveys information on the magnitude of the correlation among the original variables. Furthermore, it can be shown that for the regression of $r(Y|X)$ on $r(X|Y)$ the usual moment estimate of the variance around the OLS line is equal to

$$\hat{\sigma}_{r(Y|X)|r(X|Y)}^2 = (1 - r^2) \frac{n - 1}{n - 2} s_y^2,$$

and then it is inversely related to the squared correlation coefficient. Hence, if we plot $r(Y|X)$ against $r(X|Y)$ both the slope and the spread of the observed points around the OLS regression line will gather information on the correlation among X and Y .

4. The proposed graphical method

Consider a scatterplot of (x_i, y_i) , and let the interdependence among the two variables be the issue. As well known, to get consistency in the interpretation we should first take into account some basic plotting guidelines such as the correct aspect ratio (Cleveland *et al.*, 1988) and few others, for which we refer the reader to the discussion in Cook and Weisberg (1999), pp. 66–68.

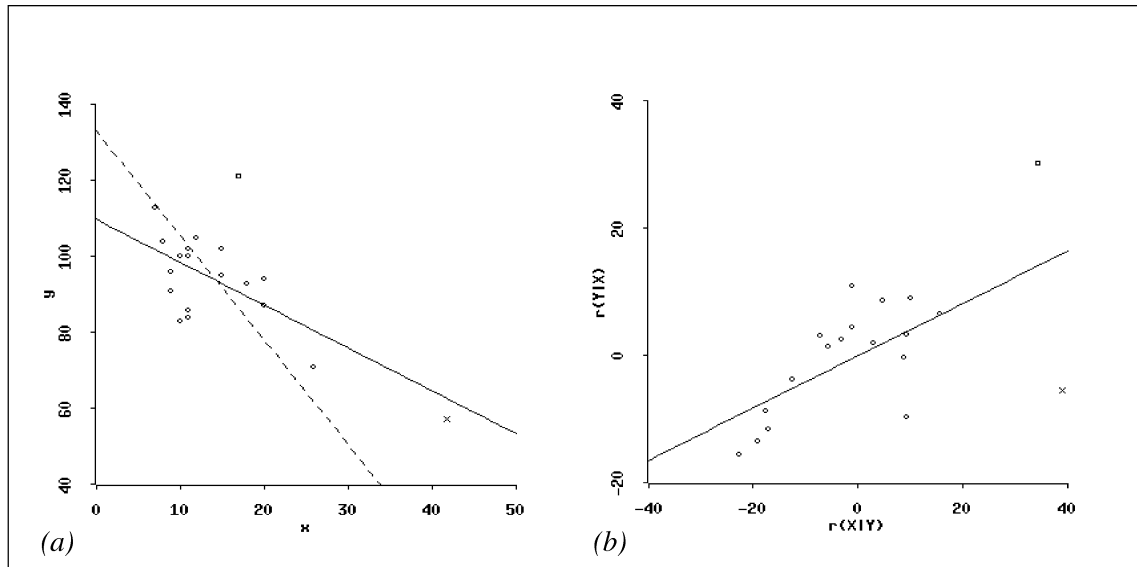
Then, to analyze the linear interdependence among the variables, we propose to consider the following graphical procedure:

- plot Y vs X , along with the two regression lines $E(Y|X)$ and $E(X|Y)$;
- plot $r(Y|X)$ vs $r(X|Y)$, along with the corresponding OLS line;
- link the two plots through a dynamic graphics device.

The first plot allows to assess graphically the sign and, to some extent, the strength of the relationship among X and Y . Increasing lines will indicate positive association and viceversa, while the width of the angle between them will stay for the strength.

The second plot will show a line passing through the origin and with slope equal to the squared correlation coefficient between X and Y . From such a graph the strength of the relationship can be visually assessed by looking at both the slope of the line and the spread of the points around it: weak correlations will be displayed by lines approximately parallel to the horizontal axis and by a large spread of points around them. As the correlation increases, the slope of the straight line (the squared correlation coefficient), will tend to the $\pi/4$ angle, while the points will tend to lie more and more closer. In the case of perfect correlation ($|r| = 1$), all the points will lie exactly on the 45° straight line.

Figure 1: The proposed graphical method applied to an artificial data set: (a) scatterplot with the two regression lines; (b) the new plot to investigate the correlation structure.



At the end, dynamically linking the plots allows detection of potentially influential observations. Points far from the others in the $r(Y|X)$ vs $r(X|Y)$ plot will be good candidates to be influential on the observed correlation coefficient. Their omission will allow direct visualization of their influence on the sign and the magnitude of the coefficient. Deletion of influential data will modify the angle among the two lines in the Y vs X plot and both the slope and the spread of the points in the $r(Y|X)$ vs $r(X|Y)$ plot.

In order to illustrate the proposed methodology, we present the discussed plots for an artificial 20 points data set. The regression lines in the scatterplot (Figure 1 (a)) highlight a certain degree of negative correlation ($r = -0.64$), while the slope of the OLS line in the $r(Y|X)$ vs $r(X|Y)$ plot (Figure 1 (b)) visually measures the strength of the relationship among the observed variables ($r^2 = 0.41$). Furthermore, on the right margin of the latter plot we note two isolated points potentially influential on both the slope of the line and the spread of the points around it, and therefore potentially influential on the observed correlation coefficient. Their omission yield $r = -0.52$.

Finally, we note that the proposed graphical procedure is particularly worth pursuing when few observations are available, as it appears not easy to evaluate by eye from the scatterplot the strength of the association. On the other hand, if a large number of observations are available, our method helps behind possible misleading overplotting effect.

References

- Bowman A. and Young S. (1996) Graphical comparison of nonparametric curves, *Applied Statistics*, 45, 83–98.
- Cleveland W.S., McGill M.E. and McGill R. (1988) The shape parameter of a two-variable graph, *Journal of the American Statistical Association*, 83, 289–300.
- Cook R. and Weisberg S. (1999) *Applied Regression Including Computing and Graphics*, Wiley, New York.
- Kendall M. and Stuart A. (1967) *The Advanced Theory of Statistics, Vol 2, Second Edition*, Charles Griffin, London.