Coefficients of Correlation, Alienation and Determination

Hervé Abdi[1]

1 Overview

The coefficient of correlation evaluates the similarity of two sets of measurements (*i.e.,* two dependent variables) obtained on the same observations. The coefficient of correlation indicates the amount of information common to two variables. This coefficient takes values between −1 and +1 (inclusive). A value of 0 indicates that the two series of measurement have nothing in common. A value of +1 says that the two series of measurements are measuring the same thing. A value of −1 says that the two measurements are measuring the same thing but one measurement varies inversely to the other.

The squared correlation gives the proportion of common variance between two variables and is also called the *coefficient of determination.* Subtracting the coefficient of determination from the unity gives the proportion of variance not shared between two variables, a quantity also called the *coefficient of alienation.*

The coefficient of correlation measures only the *linear* relationship between two variables, and its value is very sensitive to

[1]In: Neil Salkind (Ed.) (2007). *Encyclopedia of Measurement and Statistics.* Thousand Oaks (CA): Sage.
Address correspondence to: Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
E-mail: herve@utdallas.edu http://www.utd.edu/~herve

outliers. Its significance can be tested with an F or a t test. The coefficient of correlation always overestimates the intensity of the correlation in the population and needs to be "corrected" in order to provide a better estimation. The corrected value is called "shrunken" or "adjusted."

2 Notations and definition

We have S observations, and for each observation we have two measurements denoted W and Y with respective means M_W and M_Y. For each observation, we define the cross-product as the product of the deviations of each variable to its mean. The sum of these cross-products, denoted SCP_{WY}, is computed as:

$$SCP_{WY} = \sum_{s}^{S} (W_s - M_W)(Y_s - M_Y) \; . \tag{1}$$

The sum of the cross-products reflects the association between the variables. When the deviations tend to have the same sign, they indicate a positive relationship, when they tend to have different signs, they indicate a negative relationship. The average value of the SCP_{WY} is called the covariance [just as the variance, the covariance can be computed by dividing by S or $(S-1)$]:

$$\text{cov}_{WY} = \frac{SCP}{\text{Number of Observations}} = \frac{SCP}{S} \; . \tag{2}$$

The covariance reflects the association between the variables but it is expressed in the original units of measurement. In order to eliminate them, the covariance is normalized by division by the standard deviation of each variable. This defines the coefficient of correlation denoted $r_{W.Y}$ which is equal to

$$r_{W.Y} = \frac{\text{cov}_{WY}}{\sigma_W \sigma_Y} \; . \tag{3}$$

Rewriting the previous formula, gives a more practical formula:

$$r_{W.Y} = \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}} \; . \tag{4}$$

3 An example: Correlation computation

We illustrate the computation for the coefficient of correlation with the following data, describing the values of W and Y for $S = 6$ subjects:

$$W_1 = 1 \quad W_2 = 3 \quad W_3 = 4 \quad W_4 = 4 \quad W_5 = 5 \quad W_6 = 7$$

$$Y_1 = 16 \quad Y_2 = 10 \quad Y_3 = 12 \quad Y_4 = 4 \quad Y_5 = 8 \quad Y_6 = 10$$

Step 1: Computing the sum of the cross-products
First compute the means of W and Y:

$$M_W = \frac{1}{S}\sum_{s=1}^{S} W_s = \frac{24}{6} = 4 \text{ and } M_Y = \frac{1}{S}\sum_{s=1}^{S} Y_s = \frac{60}{6} = 10 .$$

The sum of the cross-products is then equal to

$$\begin{aligned} SCP_{YW} &= \sum_s (Y_s - M_Y)(W_s - M_W) \\ &= (16-10)(1-4) + (10-10)(3-4) + (12-10)(4-4) \\ &\quad + (4-10)(4-4) + (8-10)(5-4) + (10-10)(7-4) \\ &= (6\times -3) + (0\times -1) + (2\times 0) + (-6\times 0) + (-2\times 1) + (0\times 3) \\ &= -18+0+0+0-2+0 \\ &= -20 . \end{aligned} \tag{5}$$

The sum of squares of W_s is obtained as

$$\begin{aligned} SS_W &= \sum_{s=1}^{S} (W_s - M_W)^2 \\ &= (1-4)^2 + (3-4)^2 + (4-4)^2 + (4-4)^2 + (5-4)^2 + (7-4)^2 \\ &= (-3)^2 + (-1)^2 + 0^2 + 0^2 + 1^2 + 3^2 \\ &= 9+1+0+0+1+9 \\ &= 20 . \end{aligned} \tag{6}$$

The sum of squares of Y_s is

$$\begin{aligned} SS_Y &= \sum_{s=1}^{S} (Y_s - M_Y)^2 \\ &= (16-10)^2 + (10-10)^2 + (12-10)^2 + (4-10)^2 \\ &\quad + (8-10)^2 + (10-10)^2 \\ &= 6^2 + 0^2 + 2^2 + (-6)^2 + (-2)^2 + 0^2 \\ &= 36 + 0 + 4 + 36 + 4 + 0 \\ &= 80 . \end{aligned} \tag{7}$$

Step 3: Computing $r_{W.Y}$
The coefficient of correlation between W and Y is equal to

$$\begin{aligned} r_{W.Y} &= \frac{\sum_s (Y_s - M_Y)(W_s - M_W)}{\sqrt{SS_Y \times SS_W}} \\ &= \frac{-20}{\sqrt{80 \times 20}} = \frac{-20}{\sqrt{1600}} = -\frac{20}{40} \\ &= -.5 . \end{aligned} \tag{8}$$

We can interpret this value of $r = -.5$ as an indication of a negative linear relationship between W and Y.

4 Some Properties of the coefficient of correlation

The coefficient of correlation is a number *without unit*. This occurs because dividing the units of the numerator by the same units in the denominator eliminates these units. Hence, the coefficient of correlation can be used to compare different studies performed different variables. The magnitude of the coefficient of correlation is always smaller than or equal to 1. This happens because the numerator of the coefficient of correlation (see Equation 4) is

always smaller than or equal to its denominator (this property follows from the Cauchy-Schwartz inequality). A coefficient of correlation equals to +1 or −1 indicates that a plot of the observations will show that they are positioned on a line.

The squared coefficient of correlation gives the *proportion of common variance* between two variables. It is also called the *coefficient of determination*. In our example, the coefficient of determination is equal to $r^2_{W.Y} = .25$. The proportion of variance not shared between the variables is called the *coefficient of alienation*, for our example, it is equal to $1 - r^2_{W.Y} = .75$.

5 Interpreting correlation

5.1 Linear and nonlinear relationship

The coefficient of correlation measures only *linear* relationship between two variables and will miss *non-linear* relationships. For example, Figure 1 displays a perfect nonlinear relationship between two variables (*i.e.,* the data show a U-shaped relationship with Y being proportional to the square of W), but the coefficient of correlation is equal to 0.

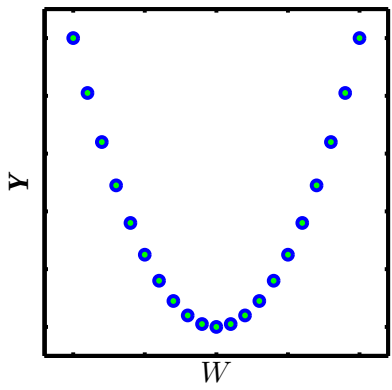

Figure 1: A perfect nonlinear relationship with a 0 correlation.

5.2 The effect of outliers

Observations far from the center of the distribution contribute a lot to the sum of the cross-products. At the extreme, in fact, as illustrated in Figure 2, one extremely deviant observation (often called an “outlier") can dramatically influence the value of r.

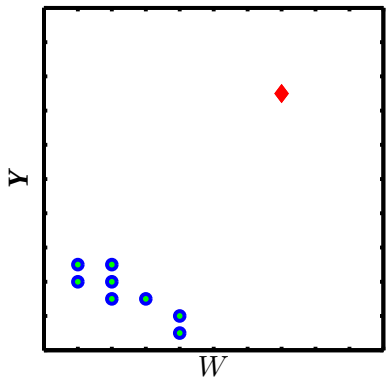

Figure 2: The dangerous effect of outliers on the value of the coefficient of correlation. The correlation of the set of points represented by the circles is equal to $-.87$, when the point represented by the diamond is added to the set, the correlation is now equal to $+.61$. This shows that an outlier can completely determine the value of the coefficient of correlation.

5.3 Geometric interpretation: The coefficient of correlation is a cosine

Each set of observations can also be seen as a *vector* in an S dimensional space (one dimension per observation). Within this framework, the correlation is equal to the *cosine* of the angle between the two vectors after they have been centered by subtracting their respective mean. For example, a coefficient of correlation of $r = -.50$ corresponds to a 150-degree angle. A coefficient of correlation of 0 corresponds to a right angle and therefore two uncorrelated variables are called *orthogonal* (which is derived from the Greek word for “right-angle”).

5.4 Correlation and causation

The fact that two variables are correlated does not mean that one variable causes the other one: *correlation is not causation*. For example: in France, the number of Catholic churches, as well as the number of schools, in a city are highly correlated with the number of cirrhosis of the liver, the number of teenage pregnancies, and the number of violent deaths. Does that mean that churches and schools are sources of vice and that newborns are murderers? Here, in fact, the observed correlation is due to a third variable, namely the size of the cites: the larger a city, the larger the number of churches, schools and alcoholics, etc. In this example, the correlation between number of churches/schools and alcoholics is called a *spurious* correlation because it reflects only their mutual correlation with a third variable (*i.e.,* size of the city).

6 Testing the significance of r

A null hypothesis test for r can be performed using an F statistic obtained as:

$$F = \frac{r^2}{1-r^2} \times (S-2)\ . \tag{9}$$

When the null hypothesis is true (and when the normality assumption holds), this statistic is distributed as a Fisher's F with $\nu_1 = 1$ and $\nu_2 = S-2$ degrees of freedom. An equivalent test can be performed using $t = \sqrt{F}$, which is distributed, under H_0 as a Student's distribution with $\nu = S-2$ degrees of freedom.

For our example, we find that

$$F = \frac{.25}{1-.25} \times (6-2) = \frac{.25}{.75} \times 4 = \frac{1}{3} \times 4 = \frac{4}{3} = 1.33\ .$$

The probability of finding such a value under H_0 is found using an F distribution with $\nu_1 = 1$ and $\nu_2 = 3$, and is equal to $p \approx .31$. Such a value does not lead to rejecting H_0.

7 Estimating the population correlation: shrunken and adjusted *r*

The coefficient of correlation is a *descriptive* statistic which always overestimates the population correlation. This problem is similar to the problem of the estimation of the variance of a population from a sample. In order to obtain a better estimate of the population, the value r needs to be corrected. The corrected value of r goes under different names: corrected r, shrunken r, or adjusted r (there are some subtle differences between these different appellations, but we will ignore them here) and we denote it by $\tilde{r}^2$. There are several correction formulas available, the one most often used estimates the value of the population correlation as

$$\tilde{r}^2 = 1 - \left[\left(1 - r^2\right)\left(\frac{S-1}{S-2}\right)\right] . \tag{10}$$

For our example, this gives:

$$\tilde{r}^2 = 1 - \left[\left(1 - r^2\right)\left(\frac{S-1}{S-2}\right)\right] = 1 - \left[(1 - .25) \times \frac{5}{4}\right] = 1 - \left[.75 \times \frac{5}{4}\right] = 0.06 .$$

With this formula, we find that the estimation of the population correlation drops from $r = .-50$ to $\tilde{r} = -\sqrt{\tilde{r}^2} = -\sqrt{.06} = -.24$.

References

[1] Cohen, J., & Cohen, P. (1983) *Applied multiple regression / correlation analysis for the social sciences*. Hillsdale (NJ): Erlbaum.

[2] Darlington, R.B. (19 90). *Regression and linear models*. New York: McGraw-Hill.

[3] Pedhazur, E.J. (1997) *Multiple regression in behavioral research*. New York: Harcourt-Brace.