Multiple Correlation Coefficient

Hervé Abdi[1]

1 Overview

The *multiple* correlation coefficient generalizes the standard coefficient of correlation. It is used in multiple regression analysis to assess the quality of the prediction of the dependent variable. It corresponds to the squared correlation between the predicted and the actual values of the dependent variable. It can also be interpreted as the proportion of the variance of the dependent variable explained by the independent variables. When the independent variables (used for predicting the dependent variable) are pairwise orthogonal, the multiple correlation coefficient is equal to the sum of the squared coefficients of correlation between each independent variable and the dependent variable. This relation does not hold when the independent variables are not orthogonal. The significance of a multiple coefficient of correlation can be assessed with an F ratio. The magnitude of the multiple coefficient of correlation tends to overestimate the magnitude of the population correlation, but it is possible to correct for this overestimation. Strictly speaking we should refer to this coefficient as the *squared* multiple correlation coefficient, but current usage seems to ignore the

[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

adjective "squared," probably because mostly its squared value is considered.

2 Multiple Regression framework

In linear multiple regression analysis, the goal is to predict, knowing the measurements collected on N subjects, a dependent variable Y from a set of J independent variables denoted

$$\{X_1, \ldots, X_j, \ldots, X_J\} . \tag{1}$$

We denote by $\mathbf{X}$ the $N \times (J+1)$ augmented matrix collecting the data for the independent variables (this matrix is called augmented because the first column is composed only of ones), and by $\mathbf{y}$ the $N \times 1$ vector of observations for the dependent variable. These two matrices have the following structure.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,j} & \cdots & x_{n,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,j} & \cdots & x_{N,J} \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{bmatrix} \tag{2}$$

The predicted values of the dependent variable $\widehat{Y}$ are collected in a vector denoted $\widehat{\mathbf{y}}$ and are obtained as:

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} \quad \text{with} \quad \mathbf{b} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} . \tag{3}$$

The regression sum of squares is obtained as

$$SS_{\text{regression}} = \mathbf{b}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y} - \frac{1}{N}(\mathbf{1}^{\mathsf{T}}\mathbf{y})^2 \tag{4}$$

(with $\mathbf{1}^{\mathsf{T}}$ being a row vector of 1's conformable with $\mathbf{y}$).

The total sum of squares is obtained as

$$SS_{\text{total}} = \mathbf{y}^{\mathsf{T}}\mathbf{y} - \frac{1}{N}(\mathbf{1}^{\mathsf{T}}\mathbf{y})^2 . \tag{5}$$

The residual (or error) sum of squares is obtained as

$$SS_{\text{error}} = \mathbf{y}^{\mathsf{T}}\mathbf{y} - \mathbf{b}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y}\,. \tag{6}$$

The quality of the prediction is evaluated by computing the multiple coefficient of correlation denoted $R^2_{Y.1,\ldots,J}$. This coefficient is equal to the squared coefficient of correlation between the dependent variable (Y) and the predicted dependent variable ($\widehat{Y}$).

An alternative way of computing the multiple coefficient of correlation is to divide the regression sum of squares by the total sum of squares. This shows that $R^2_{Y.1,\ldots,J}$ can also be interpreted as the proportion of variance of the dependent variable explained by the independent variables. With this interpretation, the multiple coefficient of correlation is computed as

$$R^2_{Y.1,\ldots,J} = \frac{SS_{\text{regression}}}{SS_{\text{regression}} + SS_{\text{error}}} = \frac{SS_{\text{regression}}}{SS_{\text{total}}}\,. \tag{7}$$

2.1 Significance test

In order to assess the significance of a given $R^2_{Y.1,\ldots,J}$, we can compute an F ratio as

$$F = \frac{R^2_{Y.1,\ldots,J}}{1 - R^2_{Y.1,\ldots,J}} \times \frac{N-J-1}{J}\,. \tag{8}$$

Under the usual assumptions of normality of the error and of independence of the error and the scores, this F ratio is distributed under the null hypothesis as a Fisher distribution with $\nu_1 = J$ and $\nu_2 = N-J-1$ degrees of freedom.

2.2 Estimating the population correlation: shrunken and adjusted R

Just like its bivariate counterpart r, the multiple 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.

Table 1: A set of data. The dependent variable Y is to be predicted from two orthogonal predictors X_1 and X_2 (data from Abdi *et al.*, 2002). These data are the results of an hypothetical experiment on retroactive interference and learning. Y is the number of sentences remembered from a set of sentences learned, X_1 is the number of learning trials, and X_2 is the number of interpolated lists learned.

Number of learning trials (X)	Number of interpolated lists (T)		
	2	**4**	**8**
2	35	21	6
	39	31	8
4	40	34	18
	52	42	26
8	61	58	46
	73	66	52

In order to obtain a better estimate of the population, the value $R^2_{Y.1,\ldots,J}$ needs to be corrected. The corrected value of $R^2_{Y.1,\ldots,J}$ 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 $\widetilde{R}^2_{Y.1,\ldots,J}$. There are several correction formulas available, the one most often used estimates the value of the population correlation as

$$\widetilde{R}^2_{Y.1,\ldots,J} = 1 - \left[\left(1 - R^2_{Y.1,\ldots,J}\right)\left(\frac{N-1}{N-J-1}\right)\right] . \tag{9}$$

3 Example 1: Multiple correlation coefficient with orthogonal predictors

When the independent variables are pairwise orthogonal, the importance of each of them in the regression is assessed by computing the squared coefficient of correlation between each of the independent variables and the dependent variable. The sum of these squared coefficients of correlation is equal to the multiple coefficient of correlation. We illustrate this case with the data from Table 1. In this example, the dependent variable (Y) is the number or sentences recalled by participants who learned a list of unrelated sentences. The first independent variable or first predictor, X_1 is the number of trials used to learn the list. It takes the values 2, 4, and 8. It is expected that recall will increase as a function of the number of trials. The second independent variable, X_2 is the number of additional interpolated lists that the participants are asked to learned. It takes the values 2, 4, and 8. As a consequence of retroactive inhibition, it is expected that recall will decrease as a function of the number of interpolated lists learned.

Using Equation 3, we found that $\widehat{Y}$ can be obtained from X_1 and X_2 as

$$\widehat{Y} = 30 + 6 \times X_1 - 4 \times X_2. \tag{10}$$

Using these data and Equations 4 and 5, we find that

$$SS_{\text{regression}} = 5824, \quad SS_{\text{total}} = 6214, \quad \text{and} \quad SS_{\text{error}} = 390\,. \tag{11}$$

This gives the following value for the multiple coefficient of correlation:

$$R^2_{Y.1,\ldots,J} = \frac{SS_{\text{regression}}}{SS_{\text{total}}} = \frac{5824}{6214} = .9372\,. \tag{12}$$

In order to decide if this value of $R^2_{Y.1,\ldots,J}$ is large enough to be considered significant, we compute an F ratio equal to

$$F = \frac{R^2_{Y.1,\ldots,J}}{1 - R^2_{Y.1,\ldots,J}} \times \frac{N-J-1}{J} = \frac{.9372}{1-.9372} \times \frac{15}{2} = 111.93\,. \tag{13}$$

Such a value of F is significant at all the usual alpha levels, and therefore we can reject the null hypothesis.

Because X_1 and X_2 are orthogonal to each other (*i.e.*, their correlation is equal to 0), the multiple coefficient of correlation is equal to the sum of the squared coefficients of correlation between the independent variables and the dependent variable:

$$R^2_{Y.1,\ldots,J} = .9372 = r^2_{Y,1} + r^2_{Y,2} = .6488 + .2884 \; . \tag{14}$$

A better estimate of the population value of the multiple coefficient of correlation can obtained as

$$\widetilde{R}^2_{Y.1,\ldots,J} = 1 - \left[\left(1 - R^2_{Y.1,\ldots,J}\right) \left(\frac{N-1}{N-J-1} \right) \right] = 1 - (1 - .9372) \frac{17}{15} = .9289 \; . \tag{15}$$

4 Example 2: Multiple correlation coefficient with non-orthogonal predictors

When the independent variables are correlated, the multiple coefficient of correlation is not equal to the sum of the squared correlation coefficients between the dependent variable and the independent variables. In fact, such a strategy would *overestimate* the contribution of each variable because the variance that they share would be counted several times.

For example, consider the data given in Table 2 where the dependent variable is to be predicted from the independent variables X_1 and X_2. The prediction of the dependent variable (using Equation 3) is found to be equal to

$$\widehat{Y} = 1.67 + X_1 + 9.50 X_2 \; ; \tag{16}$$

this gives a multiple coefficient of correlation of $R^2_{Y.1,\ldots,J} = .9866$. The coefficient of correlation between X_1 and X_2 is equal to $r_{X_1.X_2} = .7500$, between X_1 and Y is equal to $r_{Y.1} = .8028$, and between X_2 and Y is equal to $r_{Y.2} = .9890$. It can easily be checked that the

Table 2: A set of data. The dependent variable Y is to be predicted from two correlated (*i.e.*, non-orthogonal) predictors: X_1 and X_2 (data from Abdi *et al.*, 2002). Y is the number of digits a child can remember for a short time (the "memory span"), X_1 is the age of the child, and X_2 is the speech rate of the child (how many words the child can pronounce in a given time). Six children were tested.

Y (Memory span)	14	23	30	50	39	67
X_1 (age)	4	4	7	7	10	10
X_2 (Speech rate)	1	2	2	4	3	6

multiple coefficient of correlation is not equal to the sum of the squared coefficients of correlation between the independent variables and the dependent variables:

$$R^2_{Y.1,\ldots,J} = .9866 \neq r^2_{Y.1} + r^2_{Y.2} = .665 + .9780 = 1.6225\,. \tag{17}$$

Using the data from Table 2 along with Equations 4 and 5, we find that

$$SS_{\text{regression}} = 1822.00, \quad SS_{\text{total}} = 1846.83, \quad \text{and} \quad SS_{\text{error}} = 24.83\,. \tag{18}$$

This gives the following value for the multiple coefficient of correlation:

$$R^2_{Y.1,\ldots,J} = \frac{SS_{\text{regression}}}{SS_{\text{total}}} = \frac{1822.00}{1846.83} = .9866\,. \tag{19}$$

In order to decide if this value of $R^2_{Y.1,\ldots,J}$ is large enough to be considered significant, we compute an F ratio equal to

$$F = \frac{R^2_{Y.1,\ldots,J}}{1 - R^2_{Y.1,\ldots,J}} \times \frac{N-J-1}{J} = \frac{.9866}{1-.9866} \times \frac{3}{2} = 110.50\,. \tag{20}$$

Such a value of F is significant at all the usual alpha levels, and therefore we can reject the null hypothesis.

A better estimate of the population value of the multiple coefficient of correlation can obtained as

$$\widetilde{R}^2_{Y.1,\ldots,J} = 1 - \left[\left(1 - R^2_{Y.1,\ldots,J}\right)\left(\frac{N-1}{N-J-1}\right)\right] = 1 - (1-.9866)\frac{5}{2} = .9776\,. \tag{21}$$

References

[1] Abdi, H., Dowling, W.J., Valentin, D., Edelman, B., & Posamentier M. (2002). *Experimental Design and research methods. Unpublished manuscript*. Richardson: The University of Texas at Dallas, Program in Cognition.

[2] Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd edition). Hillsdale (NJ): Erlbaum.

[3] Darlington, R.B. (1990). *Regression and linear models*. NewYork: McGraw-Hill.

[4] Pedhazur, E.J. (1997). *Multiple regression in behavioral research*. (3rd edition) New York: Holt, Rinehart and Winston, Inc.