Multiple correlation

From Wikipedia, the free encyclopedia

In statistics, the coefficient of **multiple correlation** is a measure of how well a given variable can be predicted using a linear function of a set of other variables. It is the correlation between the variable's values and the best predictions that can be computed linearly from the predictive variables.[1]

The coefficient of multiple correlation takes values between 0 and 1; a higher value indicates a better predictability of the dependent variable from the independent variables, with a value of 1 indicating that the predictions are exactly correct and a value of 0 indicating that no linear combination of the independent variables is a better predictor than is the fixed mean of the dependent variable.[2]

The coefficient of multiple correlation is computed as the square root of the coefficient of determination, but under the particular assumptions that an intercept is included and that the best possible linear predictors are used, whereas the coefficient of determination is defined for more general cases, including those of nonlinear prediction and those in which the predicted values have not been derived from a model-fitting procedure.

Contents

Definition

The coefficient of multiple correlation, denoted R, is a scalar that is defined as the Pearson correlation coefficient between the predicted and the actual values of the dependent variable in a linear regression model that includes an intercept.

Computation

The square of the coefficient of multiple correlation can be computed using the vector $\mathbf{c} = \left(r_{x_1 y}, r_{x_2 y}, \ldots, r_{x_N y} \right)^\top$ of correlations $r_{x_n y}$ between the predictor variables x_n (independent variables) and the target variable y (dependent variable), and the correlation matrix R_{xx} of inter-correlations between predictor variables. It is given by

$$R^2 = \mathbf{c}^\top R_{xx}^{-1} \, \mathbf{c},$$

where $\mathbf{c}^\top$ is the transpose of $\mathbf{c}$, and R_{xx}^{-1} is the inverse of the matrix

$$R_{xx} = \begin{pmatrix} r_{x_1 x_1} & r_{x_1 x_2} & \cdots & r_{x_1 x_N} \\ r_{x_2 x_1} & \ddots & & \vdots \\ \vdots & & \ddots & \\ r_{x_N x_1} & \cdots & & r_{x_N x_N} \end{pmatrix}.$$

If all the predictor variables are uncorrelated, the matrix R_{xx} is the identity matrix and R^2 simply equals $c^\top c$, the sum of the squared correlations with the dependent variable. If the predictor variables are correlated among themselves, the inverse of the correlation matrix R_{xx} accounts for this.

The squared coefficient of multiple correlation can also be computed as the fraction of variance of the dependent variable that is explained by the independent variables, which in turn is 1 minus the unexplained fraction. The unexplained fraction can be computed as the sum of squared residuals—that is, the sum of the squares of the prediction errors—divided by the sum of the squared deviations of the values of the dependent variable from its expected value.

Properties

With more than two variables being related to each other, the value of the coefficient of multiple correlation depends on the choice of dependent variable: a regression of y on x and z will in general have a different R than will a regression of z on x and y. For example, suppose that in a particular sample the variable z is uncorrelated with both x and y, while x and y are linearly related to each other. Then a regression of z on y and x will yield an R of zero, while a regression of y on x and z will yield a strictly positive R. This follows since the correlation of y with the best predictor based on x and z is in all cases at least as large as the correlation of y with the best predictor based on x alone, and in this case with z providing no explanatory power it will be exactly as large.

References

1. Introduction to Multiple Regression (http://onlinestatbook.com/2/regression/multiple_regression.html)
2. Multiple correlation coefficient (http://mtweb.mtsu.edu/stats/regression/level3/multicorrel/multicorrcoef.htm)

- Allison, Paul D. (1998). *Multiple Regression: A Primer*. London: Sage Publications. ISBN 9780761985334
- Cohen, Jacob, et al. (2002). *Applied Multiple Regression: Correlation Analysis for the Behavioral Sciences*. ISBN 0805822232
- Crown, William H. (1998). *Statistical Models for the Social and Behavioral Sciences: Multiple Regression and Limited-Dependent Variable Models*. ISBN 0275953165
- Edwards, Allen Louis (1985). *Multiple Regression and the Analysis of Variance and Covariance*. ISBN 0716710811
- Keith, Timothy (2006). *Multiple Regression and Beyond*. Boston: Pearson Education.
- Fred N. Kerlinger, Elazar J. Pedhazur (1973). *Multiple Regression in Behavioral Research*. New York: Holt Rinehart Winston. ISBN 9780030862113
- Stanton, Jeffrey M. (2001). "Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors" (http://www.amstat.org/publications/jse/v9n3/stanton.html), *Journal of Statistics Education*, 9 (3).

Retrieved from "https://en.wikipedia.org/w/index.php?title=Multiple_correlation&oldid=701131410"

Categories: Regression analysis | Covariance and correlation