Guide to Interpreting Multivariate Analyses

1. What is "r"?

Interpreting the Correlation Coefficient R. Customarily, the degree to which two or more predictors (independent or *X* variables) are related to the dependent (*Y*) variable is expressed in the correlation coefficient *R*, which is the square root of *R-square*. In multiple regression, *R* can assume values between 0 and 1.. http://www.statsoft.com/textbook/stmulreg.html)

Interpretation of the size of a correlation

Correlation	Negative	Positive
Small	−0.3 to −0.1	0.1 to 0.3
Medium	−0.5 to −0.3	0.3 to 0.5
Large	−1.0 to −0.5	0.5 to 1.0

Several authors have offered guidelines for the interpretation of a correlation coefficient. Cohen (1988),[5] has observed, however, that all such criteria are in some ways arbitrary and should not be observed too strictly. This is because the interpretation of a correlation coefficient depends on the context and purposes. A correlation of 0.9 may be very low if one is verifying a physical law using high-quality instruments, but may be regarded as very high in the social sciences where there may be a greater contribution from complicating factors.

Along this vein, it is important to remember that "large" and "small" should not be taken as synonyms for "good" and "bad" in terms of determining that a correlation is of a certain size. For example, a correlation of 1.0 or −1.0 indicates that the two variables analyzed are equivalent modulo scaling. Scientifically, this more frequently indicates a trivial result than a profound one. For example, consider discovering a correlation of 1.0 between how many feet tall a group of people are and the number of inches from the bottom of their feet to the top of their heads.

Correlation and causality

The conventional dictum that "correlation does not imply causation" means that correlation cannot be validly used to infer a causal relationship between the variables. This dictum should not be taken to mean that correlations cannot indicate causal relations. However, the causes underlying the correlation, if any, may be indirect and unknown. Consequently, establishing a correlation between two variables is not a sufficient condition to establish a causal relationship (in either direction).

A correlation between age and height in children is fairly causally transparent, but a correlation between mood and health in people is less so. Does improved mood lead to improved health; or does good health lead to good mood; or both? Or does some other

factor underlie both? Or is it pure coincidence? In other words, a correlation can be taken as evidence for a possible causal relationship, but cannot indicate what the causal relationship, if any, might be.

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, s_x and s_y are the sample Standard deviations of X and Y and the sum is from $i = 1$ to n

http://en.wikipedia.org/wiki/Correlation_coefficient

2. What is "R^2"?

Residual Variance and R-square. The smaller the variability of the residual values around the regression line relative to the overall variability, the better is our prediction. For example, if there is no relationship between the *X* and *Y* variables, then the ratio of the residual variability of the *Y* variable to the original variance is equal to 1.0. If *X* and *Y* are perfectly related then there is no residual variance and the ratio of variance would be 0.0. In most cases, the ratio would fall somewhere between these extremes, that is, between 0.0 and 1.0. 1.0 minus this ratio is referred to as *R-square* or the *coefficient of determination*. This value is immediately interpretable in the following manner. If we have an *R-square* of 0.4 then we know that the variability of the *Y* values around the regression line is 1-0.4 times the original variance; in other words we have explained 40% of the original variability, and are left with 60% residual variability. Ideally, we would like to explain most if not all of the original variability. The *R-square* value is an indicator of how well the model fits the data (e.g., an *R-square* close to 1.0 indicates that we have accounted for almost all of the variability with the variables specified in the model). (the correlation coefficient *R* is the square root of *R-square or* R^2*)*
http://www.statsoft.com/textbook/stmulreg.html)

3. What is a regression coefficient "B"or "β"?

The regression coefficients. The regression coefficients (or *B* coefficients) represent the *independent* contributions of each independent variable to the prediction of the dependent variable. To interpret the direction of the relationship between variables, one looks at the signs (plus or minus) of the regression or *B* coefficients. If a *B* coefficient is positive, then the relationship of this variable with the dependent variable is positive (e.g., the greater the IQ the better the grade point average); if the *B* coefficient is negative then the relationship is negative (e.g., the lower the class size the better the average test scores). Of course, if the *B* coefficient is

equal to 0 then there is no relationship between the variables.
http://www.statsoft.com/textbook/stmulreg.html)

“B”or “β”? Regression coefficients come from regression equations run on variables. A regression run on original, unstandardized variables produces unstandarized coefficients while a regression run on standardized variables produces standardized coefficients. The standardization is conducted to answer the question of which of the independent variables have a greater effect on the dependent variable in a multiple regression analysis when the variables are measures in different units (IQ score in pts, income in dollars, blood lead in ug/dL, family size in number of individuals, education in number of years)

The standardized regression coefficients, then, represent the change in terms of standard deviations in the dependent variable that result from a change of one standard deviation in an independent variable. Standardized coefficients are labeled " **β** " while the ordinary unstandarized coefficients are labeled "B"

Advocates of standardized regression coefficients point out that the coefficients are the same regardless of an independent variable's underlying scale of units. They also suggest that this removes the problem of comparing, for example, years with kilograms since each regression coefficient represents the change in response per standard unit (one SD) change in a predictor.

However, critics of standardized regression coefficients argue that this is illusory: there is no reason why a change of one SD in one predictor should be equivalent to a change of one SD in another predictor. Some variables are easy to change--the amount of time watching television, for example. Others are more difficult--weight or cholesterol level. Others are impossible--height or age.

http://en.wikipedia.org/wiki/Standardized_coefficient