Multiple Correlation and Mulitvariate Regression

Examining the strength, direction and variance explained in a set of variables
Making predictions about the value of one variable based on a set of variables

Rarely in the real world does just one factor contribute to a particular outcome; it is more frequently a combination of factors. Multiple correlation and multivariate regression makes it possible to look at a set of variables and their impact on an outcome, both as a combined set and as individual predictors. Two or more variables in combination make better predictions than one alone.

Multiple Correlation

Multiple correlation determines how much variation in the dependent variable is explained by a set of independent variables. When adding additional independent variables to the model, the value of the R^2 may increase but it cannot decrease.

The adjusted R^2 takes into account how little variation an additional variable in the model is contributing to the explained variation in the dependent variable. This adjusted R^2 will decrease with each additional variable if those variables are adding little to the explained variance. Large differences between the adjusted R^2 and the R^2 suggest that there are many superfluous variables in the model.

Multivariate Regression

Multivariate regressions include two or more independent variables (I/R or dummy variables) in the regression analysis. When doing so, the coefficient estimates (b) for each of the independent variables are determined by holding constant (controlling for) the effects of all other variables in the model. This allows each b to serve as the best possible predictor for that particular independent variable, independent of the effects of the other variables in the model.

In order to determine the "true" effects of each of your independent variables on the dependent variable, it is necessary to include all relevant variables within your regression model – e.g. those from alternative hypotheses or variables shown to have an impact from previous research.

The F test determines the significance of the model, or set of variables.

Making predictions

When making predictions, one uses the same regression equation as in bivariate regressions but includes the b coefficient values and independent variable values for each variable in the model. The resulting Y′ is the value of the dependent variable for an individual who has the characteristics included in the equation.

MULTIPLE REGRESSION EQUATION

$$Y' = a + b_1X_1 + b_2X_2 + \ldots b_kX_k$$

where
- Y' = predicted value of the criterion, or dependent, variable
- a = y-intercept (referred to as the constant), which is derived from the means of the observed criterion and predictor variables
- X_1 = first variable entered into the equation
- b_1 = slope associated with the first variable entered into the equation (i.e., the amount of change in Y for each unit increase in X_1^*)
- X_2 = second variable entered into the equation
- b_2 = slope associated with the second variable entered into the equation (i.e., the amount of change in Y for each unit increase in X_2^*)
- X_k = last variable entered into the equation
- b_1 = slope associated with the last variable entered into the equation (i.e., the amount of change in Y for each unit increase in X_k^*)

*The effects of the other predictors are held aside, or held constant.

Log Transformations

Log transformations of dependent variables make it possible deal with outliers and the effects they have on estimates. In essence, the transformation makes the distribution more normal.

To interpret b coefficients in log dependent regressions, convert the b coefficient to a percentage by multiplying it by 100. This percentage reflects the percent increase or decrease for each unit increase in the independent variable (for I/R variables) or the percent more or less when comparing different groups (for dummy variables). By comparing the percentages, one can determine which coefficient/variable has the greatest impact on the dependent variable.

Language for I/R variables

For every unit increase in *[X/independent variable],* the value of *[Y/dependent variable]* is predicted to *[increase or decrease]* by *[b%].*

Examples:

- For each additional hour worked, a child's Behavioral Problems Index score is predicted to increase by 3%.
- For each additional hour of TV watched, a child's reading comprehension score is predicted to decrease by 1.5%.

Language for dummy variables

[Those in the included group] are predicted to have *[Y/dependent variable]* that is *[b%] [more or less]* than *[those in the excluded group].*

Examples:

- Those with college degrees are predicted to have incomes that are 27% higher than those who do not have college degrees.
- Those who take public transportation to work are predicted to score 13% lower on a stress test than those who do not commute via public transportation.