Correlation Analysis and Model Building

Correlations between the response variable and potential predictors can be useful by suggesting variables that should be included or excluded from model building. Often, modelers have many predictors, and thus, a very large number of possible models to explore. Predictors with a weak or no relationship with the response variable might sometimes be excluded. Typically, the decision to throw out a variable is based on multivariable analyses. However, if the modeler has far more predictors than can be used and variable reduction becomes necessary (often under time pressure), predictors with weak or no correlation with the response variable are good candidates for exclusion. Part of correlation analysis involves visually assessing associations between variables by looking at scatter plots. When these plots reveal patterns in the data, such as curvilinear relationships, a modeler might need to build additional terms into the model, such as polynomials. Another reason to create scatter plots is to assess the linear relationship between pairs of predictor variables. When predictors are highly correlated, they provide redundant information. Multicollinearity (strong correlations among sets of predictors) can destabilize parameter estimates and degrade the ability of model selection routines, such as stepwise selection, to select good variables. Correlation analysis is one of several ways to address collinearity prior to model building.