StatNews #65

Diagnosing Multicollinearity
October 2004

Multiple regression is a versatile statistical tool and is useful in many fields. It involves determining how a set of predictor variables are related to a response variable. Like all statistical analyses, however, certain assumptions must be met. Multiple regression assumes that each predictor variable represents at least some unique information. This assumption is sometimes violated when there is substantial redundancy among the predictor variables.

This "substantial redundancy" is known as multicollinearity, which occurs when the predictor variables, as a set, contain less total information than would appear. Multicollinearity can have two effects on the regression analysis. First, the regression parameters will be unstable from sample to sample because the standard errors of the regression parameters are very large. Second, the interpretation of the regression parameters as the effect of one predictor while holding the other predictors constant is not very informative. This is because if predictors are highly correlated, holding one predictor constant immediately limits the extent to which the other predictors can vary.

While multicollinearity can be a serious problem, this is uncommon. Using appropriate methods for diagnosing multicollinearity is important both in the rare times when it causes problems and more importantly, to ease concerns when it does not.

There are three commonly used methods for diagnosing multicollinearity, but one method is greatly superior to the others and should be used to ensure accurate diagnosis. One method examines the correlation matrix of the predictor variables. High correlations between predictor variables intuitively suggest there might be multicollinearity, but this method of detection is not helpful for two reasons. First, while multicollinearity can possibly occur pair-wise, it is much more likely to involve more than two predictor variables. Second, there is no suitable diagnostic criterion for correlations to indicate "how correlated is too correlated. A more elaborate method uses Variance Inflation Factors (VIFs) to detect multicollinearity. These factors are obtained by regressing each predictor on all the other predictors and estimating an R-square value for each. While VIFs come closer to the idea of multicollinearity, they are still unable to examine all the predictors jointly.

The best way to detect multicollinearity examines all the predictor variables together. This can be done by computing condition indices--statistics based on implementing a Principal Components Analysis (see StatNews #49). The Principal Components Analysis creates new combinations (called components) of all the information in the existing predictor variables. These components are helpful because each represents unique information, i.e., they do not overlap with one another. Once the components are created, the condition indices will be computed as ratios of the variances between two components. Condition indices larger than 30 suggest there might be multicollinearity because one component may represent little or no unique information. That is, one component may have little information to represent if the information has already been "used up" by the other components.

Condition indices can be computed in SAS in PROC REG using the COLLIN option. They are also available in SPSS when performing a regression analysis. We have a handout available with more information for researchers who are not familiar with condition indices.

References:

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons.

Kleinbaum, D.G., Kupper L.L., Muller, K.E., and Nizam, A. (1998). Applied Regression Analysis and Multivariable Methods. Pacific Grove: Duxbury Press.

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1996). Applied Linear Statistical Models. Chicago: Irwin.

Author: Jennifer Schaub