Diagnosing Multicollinearity

Statnews #65

Cornell Statistical Consulting Unit

Created October 2004. Last updated October 2020

Introduction

Multicollinearity is one potential challenge of fitting a multiple linear regression model. If a model is proposed in which a some of the predictor variables have strong linear relationships with each other, then the resulting regression coefficients can be unstable (with large standard errors) and difficult to interpret. This situation, in which correlations or strong linear relationships exist among the predictor variables, is known as multicollinearity.

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.

Diagnosing multicollinearity

Pairwise correlations

One way of diagnosing multicollinearity involves examining the correlation matrix of the predictor variables (i.e. all pairwise correlations among the predictors). High (close to 1 in absolute value) pairwise correlations may suggest multicollinearity, but multicollinearity can still exist when all pairwise correlations are close to zero. This is because linear relationships can exist among more than two predictors even when their pairwise correlations are small.

Variance inflation factors (VIFs)

Another way to detect multicollinearity is to calculate Variance Inflation Factors (VIFs). Each VIF corresopnds to a single predictor variable. For the jth predictor, the VIF represents the amount by which the variance of the jth regression coefficient is inflated due to linear relationships between that predictor and the remaining predictors. If a predictor has exactly zero correlation with every one of the other predictors, then its VIF is equal to one, representing no inflation of the variance of its regression coefficient. In all other cases, the jth VIF will be larger

than one, and hence the variance of the jth regression coefficient is inflated due to linear relationships between the jth predictor and the other predictors.

The jth VIF can also be thought of as a measure of the fit of a regression model where jth predictor is treated as the dependent variable, and the remaining predictors are treated as independent variables. The jth VIF is equal to $(1 - R_j^2)^{-1}$, where R_j^2 is the the multiple R-squared from this model. Since R_j^2 is close to 1 when the other predictors have strong linear relationships to the jth predictor, the VIF will be large when those linear relationships are strong.

Other approaches

Condition indices are another measure of collinearity. These statistic can be computed from 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 are always uncorrelated.

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.

Finally, multicollinearity may also be indicated when removing or adding a predictor leads to surprisingly large changes in the fitted regression coefficients, or when the sign of a regression coefficient is surprising based on substantive knowledge.

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.