Remediating Other Problems with Your Model

You know that multicollinearity is often caused by the choice of model, such as when two highly correlated predictor variables are used in the regression equation. In these situations, you can lessen the impact of multicollinearity by respecifying the regression equation.

One approach to model respecification is to redefine the predictor variables. For example, if X_1, X_2, and X_3 are nearly linearly dependent, it might be possible to find some function such as $X = (X_1 + X_2)/X_3$ or $X = X_1X_2X_3$ that preserves the information provided by the original predictor variables, but improves the fit of the regression model and minimizes the impact due to multicollinearity. Another widely used approach to model respecification is to eliminate redundant predictor variables. That is, if X_1, X_2 and X_3 are nearly linearly dependent, eliminating one predictor variable (say x_3) might be helpful in combating multicollinearity.

Variable elimination is often a highly effective technique. However, it might not provide a satisfactory solution if the predictor variables dropped from the model have significant explanatory power relative to the response Y. You know that ordinary least squares estimators provide unbiased estimates of parameters and that the estimates have minimum variance among all unbiased estimators. But in the presence of multicollinearity, the minimum variance of the parameter estimates might be unacceptably large.

One way to alleviate this problem is to drop the requirement that the estimator of β (the regression coefficient) be unbiased. It might be better to use a biased estimator that has a smaller variance. You can use biased regression techniques such as ridge regression and principal component regression to obtain biased estimators of regression coefficients.

In ridge regression, you reduce the variances of the parameter estimates by considering a matrix, X'X + *k*I, where *k* is a small positive quantity referred to as a shrinkage parameter. The choice of *k* is a compromise between decreasing variance and increasing bias. You can compute ridge regression estimates for a set of values of *k* starting with *k* = 0 (the unbiased estimate). A plot of the coefficients against *k* (also called ridge traces) enables you to choose the value of *k* where most of the changes in the parameter estimates are realized. To perform ridge regression, you use the RIDGE= option in the MODEL statement of PROC REG to specify the range of values of *k*. To obtain the ridge traces, you use the RIDGEPLOT option in the PLOT statement. Because the selection of *k* is subjective, you should select a value based on your subject matter knowledge. The parameter estimates and other statistics corresponding to the chosen *k* are the results of the ridge regression model fit to the data.

Although statisticians have proposed more objective procedures for dealing with problems related to multicollinearity, none are widely accepted. Principal component regression is another biased regression technique. This approach combats multicollinearity by using less than the full set of principal components in the model. Instead of dropping individual variables from the model, you drop linear combinations of independent variables. Each of these linear combinations of independent variables is a principal component. The first principal component explains the largest variance of the original variables (subject to a scaling constant). The second component has the second largest variance, and so on.

To compute parameter estimates using all but the last *m* principal components, you use the PCOMIT= option in the MODEL statement in PROC REG. The value *m* is one of the values specified in the list. The principal component variables are jointly uncorrelated, and therefore, eliminate the multicollinearity problem. The interpretation of the resulting model might be difficult.

Finally, as you learned in lesson 1, in polynomial regression models, you can sometimes overcome the effects of multicollinearity by centering the independent variables. You know that an influential observation singlehandedly exerts influence on the slope of a regression line and affects model statistics. So how should you remediate influential observations? Because these observations can be highly informative, you should not automatically discard a data point without justification. Instead, you examine the data and determine why the observation is influential. If the data is an error, you might be able to correct it. You can delete an influential observation if it cannot be corrected, taking care to document the situation. If appropriate, you can take other corrective actions, such as redefining the model or transforming the data. It can be useful to perform a sensitivity analysis and report the results from different scenarios. This involves analyzing the data both with and without the influential observations and evaluating how the results are affected by the inclusion and exclusion of the influential observations.

Finally, you can limit the influence of outliers by performing robust regression analysis using PROC ROBUSTREG. The main purpose of robust regression is to detect outliers and provide resistant, that is, stable, results in the presence of outliers.

Close