Print

Using an Effective Modeling Cycle

Now that you have a solid understanding of developing good regression models for your data, let's take a moment to review what you know about effective modeling.

First, you want to get to know your data by performing preliminary analyses. You should plot your data, calculate descriptive statistics, and perform correlation analysis.

Second, it's a good idea to check for collinearity before using any automated model selection techniques. This step can include the use of the correlation analyses and VIF statistics.

Third, you use PROC REG or PROC GLMSELECT to identify some good candidate models. No perfect model exists, so find the best, or simply, the most useful one. To narrow your choices to a few good models, you can use all-possible regressions if you have few candidate predictors, or stepwise selection methods when you consider many predictors.

Fourth is verifying model assumptions and searching for possible influential observations. You need to check and validate your assumptions by creating plots of residuals, and graphs of the residuals versus predicted values and predictors. To detect influential observations, you examine the RSTUDENT residuals, Cook's D statistic, DFFITS, and DFBETAS statistics.

Fifth, you need to revise your model if needed. If Steps 3 and 4 indicate the need for model revision, generate a new model by returning to these two steps. The last step is prediction testing. You should try to validate your model with a holdout data set not used to build the model to see whether it generalizes well to new data sets. You'll learn about prediction testing, and predictive modeling in general, in the next lesson, when we move from inference to prediction.

Close