Selecting Variables for Models

After gaining insights from an initial exploration of the data, the next step is to identify candidate models by selecting variables. As the number of predictors increases, and when some of those predictors have nonlinear relationships with the response, variable selection becomes more challenging. For example, it might be necessary to determine the appropriate degree of a polynomial term. Deciding on the best possible candidate models for your needs might seem overwhelming. But don't worry: you have several tools to help you select the best variables and identify good models.

First, you can use your subject matter knowledge. Using only subject matter knowledge would be a subjective approach, but your knowledge can be helpful in combination with other tools.

You can also use the information that you gather from initial data exploration. PROC GLMSELECT offers you a choice of automatic variable selection methods, which are often called model selection methods. Sequential selection techniques include forward, backward, and stepwise selection. At each step, candidate variables are evaluated for inclusion in or exclusion from the model. The forward selection method begins with only the intercept and then, at each step, adds the effect that most improves the fit. The process ends when no significant improvement can be obtained by adding any effect. The backward elimination method starts from the full model including all independent effects. Then the effects are deleted one by one until a stopping condition is satisfied. At each step, the effect that has the smallest contribution to the model is deleted. The stepwise selection method is a modification of the forward selection method. Unlike forward selection, in stepwise selection, effects already in the model do not necessarily remain in the model.

In each of these three approaches, the decision to include or exclude variables at each step is based on the model SBC by default. You can also specify model selection statistics other than the model SBC. Model selection statistics are discussed in more detail later in this lesson.

PROC GLMSELECT also offers other automatic selection methods. Like the forward selection method, the LAR (least angle regression) algorithm (Efron, 2006) produces a sequence of regression models. This method adds one parameter at each step, terminating at the full least-squares solution when all parameters have entered the model. LASSO (least absolute shrinkage and selection operator) selection (Tibshirani, 1996) arises from a constrained form of ordinary least squares regression where the sum of the absolute values of the regression coefficients is constrained to be smaller than a specified parameter. Adaptive LASSO selection is a modification of LASSO selection in which weights are applied to each of the parameters in forming the LASSO constraint. Elastic net selection (Zhou, 2005) is yet another automatic selection method available in PROC GLMSELECT.

PROC REG also offers additional model selection methods, such as all-possible model selection. For details about all of these selection methods, see the SAS documentation. Finally, as discussed earlier in this lesson, you can use PROC REG to generate residual plots that help to evaluate the model fit and model assumptions.
