Demo: Performing Model Selection Using PROC GLMSELECT

Filename: **st104d02.sas**

In this demonstration, we use four PROC GLMSELECT steps on the response variable SalePrice, regressing on eight predictor variables in the data set ameshousing3.

PROC GLMSELECT DATA=*SAS-data-set <options>***;**
*<label:>***MODEL** *dependent = regressors < / options>***;**
RUN;

1. Open program st104d02.sas.

   ```
   %let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
            Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

   /*st104d02.sas*/
   ods graphics on;
   proc glmselect data=STAT1.ameshousing3 plots=all;
           STEPWISEAIC: model SalePrice = &interval / selection=stepwise details=steps select=AIC;
           title "Stepwise Model Selection for SalePrice - AIC";
   run;

   proc glmselect data=STAT1.ameshousing3 plots=all;
           STEPWISEBIC: model SalePrice = &interval / selection=stepwise details=steps select=BIC;
           title "Stepwise Model Selection for SalePrice - BIC";
   run;

   proc glmselect data=STAT1.ameshousing3 plots=all;
           STEPWISEAICC: model SalePrice = &interval / selection=stepwise details=steps select=AICC;
           title "Stepwise Model Selection for SalePrice - AICC";
   run;

   proc glmselect data=STAT1.ameshousing3 plots=all;
           STEPWISESBC: model SalePrice = &interval / selection=stepwise details=steps select=SBC;
           title "Stepwise Model Selection for SalePrice - SBC";
   run;
   ```

   In each step, we request all default plots. We use STEPWISE as the selection method in the SELECTION= option, and include DETAILS=steps to obtain step information and the selection summary table. For each run, we specify a different selection criterion and use the SELECT= option: AIC, BIC, AICC, and SBC. Notice the corresponding labels in the MODEL statements. Again, this helps quickly identify what each PROC step is requesting.

2. Submit the code to compare the selected models.

3. Review the output.

   The first part of the output is from the run that used AIC as the selection criterion. In Step 0, the intercept-only model, the AIC value was approximately 6624. Recall that with information criteria, a smaller value is better. So in Step 1, Basement_Area is added, because it's the variable whose addition will most improve, or reduce, the AIC. The selection process continues to add or remove variables, making the AIC smaller each time.

   Now let's take a look at the summary table. We see the AIC value at each step, and the AIC for the final model is approximately 6141. It's interesting to see that, in addition to the intercept, all eight of the predictor variables were added into the model. Of course, this won't happen in every situation. The selection process stopped because all effects are in the final model and no variable could be removed. The AIC component of the Coefficient Panel shows larger improvements to the AIC across Steps 1 through 3 and moderate improvements across Steps 4 and 5. After about the fifth step, the AIC has roughly leveled off, and is showing small improvements at Steps 6 through 8. You can interpret this plot to mean that the last several variables added little improvement to the AIC, but they did add to the complexity of the model. So, in addition to the model with all eight variables, you might decide to consider some simpler models, for example, those at Steps 5, 6, and 7, as possible candidates.

   Next is the Criterion Panel. For AIC, AICC, and the adjusted R-square, the model with all eight variables is best. However, for SBC, the model at the sixth step is the best.

   We'll move on to the output from the second PROC GLMSELECT run, the model selection process that used BIC as the selection criterion. The Selection Summary table shows that, again, the final model is the one with all eight variables. Notice that the final BIC value, 5841, is different from the final AIC value that we saw earlier. That's because each information criterion uses a different calculation for the penalty.

   In the Coefficient Panel from this run, the Coefficient Progression plot is the same as the one for AIC, because all eight variables were added in the same order. Again, after the fifth step, there are small improvements in model fit.

Because this PROC GLMSELECT run uses BIC, it's added to the default plots in the criterion panel.

Now let's look at the results for the thrid PROC GLMSELECT, the model selection process that used AIIC as the selection criterion. Again, PROC GLMSELECT chooses the same eight-variable model. Finally, we'll look at the results for the last PROC GLMSELECT run that uses the selection criterion SBC. The Selection Summary table shows us that the selected model includes only six variables. At this point, the next candidate for entry is Lot_Area, but adding it wouldn't improve the SBC. The next candidate for removal is Bedroom_Above_Grade, but removing it would also not improve the SBC. Selection stopped at a local minimum of the SBC criterion.

The Coefficient Panel shows similarities to those seen earlier. Larger improvements in SBC can be seen across Steps 1 through 3, but smaller improvements occur over Steps 4 through 6. Like previous images, the standardized coefficients seem to stabilize after Step 4.

The Criterion Panel shows that, accounting only for the models that were viewed, the optimal fit statistics were obtained at Step 6.

Let's recap some of the model-building strategies that we've used on SalePrice. Using AIC, BIC, and AICC as selection criteria yields a model with all eight effects. Using SBC yields a model with only six effects. And recall that when we use significance level with stepwise, backward, and forward selection, specifying SLENTRY=0.05 and SLSTAY as 0.05, all three methods select the same model that contains seven effects.

By using multiple model-building strategies, you can generate a list of candidate models. So, how do you choose among the models? One option is to use a holdout data set to perform honest assessment on the models. Another option is to consult a subject-matter expert. Sometimes models with different sets of predictors than the one produced by an automated selection method can be more useful to you. For example, particular predictors might be added because they are important based on theoretical grounds. Some predictors might be included so that you can compare your work to previously published research. Other variables might be included specifically to control for effects, even if they were excluded by the model selection method. The cost of collecting data might help determine which model to go forward with. For example, if a five-predictor model explains as much variation as an eight-predictor model, but it requires much less time and money to collect the data, then it makes sense to use the simpler model. Finally, if a model doesn't meet statistical assumptions, it might need to be modified or excluded from the set of candidates.

Close