Selecting Candidate Models

This demonstration shows how to generate candidate models by using the automatic variable selection methods and model selection statistics available in PROC GLMSELECT. For the car company analysis, remember that the three predictors **Citympg**, **Hwympg**, and **FuelTank** appear to have a polynomial relationship with **Price**. The potential curvilinear relationship between **Price** and **Weight** is most likely due to increasing variability in the data. To reduce collinearity problems with the models, it is a good idea to center the variables when polynomial effects are created for modeling. When you use PROC GLMSELECT, it is also helpful to generate candidate models by using multiple selection methods and model selection statistics.

At the top of the code, we use the ODS GRAPHICS statement to reset the ODS Graphics options to their default settings. In the previous demonstration, remember that we turned on the imagemap feature, but we will not need it in this demonstration.

Let's skip over the macro code for a minute. The rest of the code consists of four instances of PROC GLMSELECT, each of which uses a different method to create candidate models. The first instance uses the backward elimination method with significance levels as the criterion. The second instance uses the forward selection method with significance levels. The third instance uses the backward elimination method with SBC. And the fourth instance uses the backward elimination method with adjusted R square. In each method, we want to create the same three centered polynomial effects (**P_City**, **P_Hwy**, and **P_Fuel**). However, to avoid repeating the same EFFECT statements in each instance of PROC GLMSELECT, we create a macro named **P_Eff** that contains the necessary EFFECT statements. Then, each PROC GLMSELECT statement simply calls the macro to pull in those EFFECT statements. The macro call precedes the MODEL statement, which specifies the polynomial effects. Notice that the variable list in the MODEL statement even includes two variables that are less likely to be used in the model, based on the exploratory analysis performed in the previous demonstration--**Luggage** and **Weight**

Also, notice that this demonstration uses the **cars2** data set. **Cars2** is based on the **cars** data set that was used in the previous demonstration. In **cars2**, the variables that we'll use to create polynomial effects have been standardized.

We are ready to run this code.

```
ods graphics / reset=all;

title 'Model Selection Cars2 Data Set';

%macro p_eff;
effect p_city = polynomial(citympg /degree=2
standardize(method=moments)=center);
effect p_hwy = polynomial(hwympg /degree=2
standardize(method=moments)=center);
effect p_fuel = polynomial(fueltank /degree=3
standardize(method=moments)=center);
%mend;

proc glmselect data=mydata.cars2 plots=criteria;
   title2 'Backward elimination with significance levels';
   %p_eff;
   model price = p_city p_hwy cylinders enginesize horsepower p_fuel
                 luggage weight / selection=backward select=sl
                 slstay=0.05 hierarchy = single;
run;

proc glmselect data=mydata.cars2;
   title2 'Forward selection with significance levels';
   %p_eff;
   model price = p_city p_hwy cylinders enginesize horsepower p_fuel
                 luggage weight / selection=forward select=sl
                 slentry=0.1 hierarchy=single;
run;

proc glmselect data=mydata.cars2;
```

```
   title2 'Backward elimination using SBC';
   %p_eff;
   model price = p_city p_hwy cylinders enginesize horsepower p_fuel
                 luggage weight / selection=backward select=sbc
                 hierarchy=single;
run;

proc glmselect data=mydata.cars2 plots=criteria;
   title2 'Backward elimination using adjusted R-square';
   %p_eff;
   model price = p_city p_hwy cylinders enginesize horsepower p_fuel
                 luggage weight / selection=backward select=adjrsq
                 hierarchy=single;
run;
title;
```

In the results from the first PROC GLMSELECT step (backward elimination with significance levels), the Backward Selection Summary table shows the order in which variables were removed from the model. Are we surprised that **Luggage** was eliminated first? No. In the Stop Details table, notice that the backward elimination process tried to remove the squared, centered **Hwympg** effect, but it couldn't. The value in Candidate Significance shows that this effect is significant enough to stay in the model.

Immediately below, the panel of criteria plots shows the changes in the model-fit statistics (AIC, AICC, SBC, and adjusted R square) during the backward elimination process. The star symbol on each graph represents the model with the best fit according to the specific statistic. The horizontal axis shows the effect that is removed from the model at each step. It is interesting that, according to the AIC, AICC, and adjusted R square criteria, the best model occurred at step 8, just before the centered **FuelTank** variable was removed. At step 8, AIC and AICC were at their lowest value, but the adjusted R square was at its highest. However, the SBC statistic selects the more parsimonious final model (after the centered **FuelTank** variable is removed) as the best model of those evaluated.

The selected model has three variables, **Hwympg**, **Hwympg^2**, and **Horsepower**. The model has an adjusted R square of 0.7082 and an SBC of 265.89. Because the MODEL statement in PROC GLMSELECT specified HIERARCHY=SINGLE, the final model is hierarchically well formulated.

Now, we move on to the results from the second PROC GLMSELECT step, which is forward selection with significance levels. The Forward Selection Summary table shows that this model includes the **Horsepower** and **s_FuelTank** effects. Wait a minute... these are completely different effects than in the previous candidate model! Actually, this is not surprising. When the variables considered for the model are highly correlated with one another, then the forward selection and backward elimination methods often generate very different models. In this model, the adjusted R square is 0.6929, and the SBC is 266.68. Based on these two statistics, this model appears to be inferior to the backward elimination model.

The next candidate model, backward elimination using SBC, includes the same effects as the first model (backward elimination using significance levels).

Finally, backward elimination using the adjusted R square selects a model that has eight predictor variables. However, the AIC and SBC values for this model are larger than for previous models. The selection summary shows that the process was stopped when the adjusted R square reached a local maximum. If the next candidate variable (**Citympg**) were removed from the model, the adjusted R square would drop. It is important to note that the stepwise selection processes do not guarantee that the model having the optimum value of the selection criterion statistic is selected as the final model.

Close