Print

Practice 1.4 (Level 1): Generating Candidate Models

Task

In this practice, you generate candidate models. In the data set **mydata.cars4**, the dependent variable is **LogPrice**. The independent variables in **mydata.cars4** are the same as the ones in the **mydata.cars** data set.

Reminder: Make sure you've defined the **mydata** library.

1. Write a PROC SGSCATTER step to generate plots of **LogPrice** versus all other predictor variables. Based on these plots, which variables appear to have a curvilinear relationship with **LogPrice**?

```
proc sgscatter data=mydata.cars4;
   plot LogPrice*(Citympg Hwympg Cylinders EngineSize
        Horsepower FuelTank Luggage Weight);
run;
```

   Based on the results, **Citympg**, **Hwympg**, **EngineSize**, and **Horsepower** might appear to have curvilinear relationships with **LogPrice**.

2. To create new scatter plots of the variables that exhibit curvature, use PROC SGSCATTER with the PBSPLINE option. Which variables might need to be squared in a regression model?

```
proc sgscatter data=mydata.cars4;
   plot LogPrice*(Citympg Hwympg EngineSize Horsepower) / pbspline;
run;
```

   Based on the results, it appears that **Citympg**, **Hwympg**, and **Horsepower** might need to be squared in the regression model. The variable **EngineSize** might need higher-order polynomial terms.

3. Use PROC GLMSELECT to create candidate models, as follows:
   - Use the EFFECT statement to create centered polynomial effects for the variables identified in step 2.
   - Use the following model selection methods to generate candidate models with **LogPrice** as the dependent variable:
     - backward elimination method using significance levels
     - stepwise selection method using AICC
     - forward selection using adjusted R-square
   - To request the selection criteria panel of plots for these model selection methods, add the PLOTS=CRITERIA option to the PROC GLMSELECT statement.

   How do the models generated by each selection method compare?

```
%macro e_poly;

effect p_city=polynomial(Citympg / degree=2
         standardize(method=moments)=center);
effect p_hwy=polynomial(Hwympg / degree=2
         standardize(method=moments)=center);
effect p_engine=polynomial(EngineSize / degree=2
         standardize(method=moments)=center);
effect p_hp=polynomial(Horsepower / degree=2
         standardize(method=moments)=center);
```

```
%mend;

proc glmselect data=mydata.cars4 plots= criteria;
   title 'Backward Elimination Using p-values';
   %e_poly;
   model LogPrice=p_City p_Hwy Cylinders p_Engine p_hp FuelTank
      Luggage Weight / selection=backward select=sl hierarchy=single;
run;

proc glmselect data=mydata.cars4 plots= criteria;
   title 'Stepwise Selection Using AICC';
   %e_poly;
   model LogPrice=p_City p_Hwy Cylinders p_Engine p_hp FuelTank
      Luggage Weight / selection=stepwise select=aicc
      hierarchy=single;
run;

proc glmselect data=mydata.cars4 plots= criteria;
   title 'Forward Selection Using Adjusted R-Squared';
   %e_poly;
   model LogPrice=p_City p_Hwy Cylinders p_Engine p_hp FuelTank
      Luggage Weight / selection=forward select=adjrsq
      hierarchy=single;
run;
```

As shown in the results, the models compare as follows:

- In the models based on backward elimination, the fit statistics AIC, AICC, and SBC decrease with each step. The adjusted R-square increases until step 5 and then decreases. The AIC, AICC, and SBC statistics select the model that is generated at step 6 to be the "best" model. The adjusted R-square selects the model at step 5.
- In the models based on stepwise selection, the adjusted R-square statistic rises initially and then begins to level off. The five-predictor model selected by AICC is the same as the model selected by AIC and the adjusted R square. However, SBC selects the two-predictor model with **Weight** and **Horsepower** only.
- In the models based on forward selection, the criteria panel of plots for the adjusted R-square forward selection indicates (with a star) the best fitting model among models with a given number of parameters.

4. Which variables appear to be appropriate for the regression model? Which model appears to be the best choice?

   The stepwise model based on AICC and the forward selection model based on adjusted R-square choose similar models with **EngineSize**, **Horsepower**, **Horsepower^2**, **FuelTank**, and **Weight** in common. The forward model adds **Cylinders**. These two models have nearly identical adjusted R-square values (0.7764 and 0.7765).

   The backward elimination model includes **EngineSize**, **Citympg**, **Citympg^2**, **Horsepower**, **Horsepower^2**, and **Weight**. This model has a larger adjusted R-square value (0.7964) and smaller information criteria values. Among the three models considered here, this model appears to be the best choice.

Hide Solution

Copyright © 2017 SAS Institute Inc., Cary, NC, USA. All rights reserved.

Close