Print

Practice: Building a Predictive Model Using PROC GLMSELECT

Use the **ameshousing3** data set to build a model that predicts the sale prices of homes in Ames, Iowa, that are 1500 square feet or below, based on various home characteristics.

1. Write a PROC GLMSELECT step that predicts the values of **SalePrice**. Partition the **stat1.ameshousing3** data set into a training data set of approximately 2/3 and a validation data set of approximately 1/3. Specify the seed 8675309. Define the **Interval** and **Categorical** macro variables as shown below, and use them to specify the inputs. Use stepwise regression as the selection method, Akaike's information criterion (AIC) to add and or remove effects, and average squared error for the validation data to select the best model. Add the REF=FIRST option in the CLASS statement. Submit the code and examine the results.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces
        Season_Sold Garage_Type_2 Foundation_2 Heating_QC
        Masonry_Veneer Lot_Shape_2 Central_Air;
```

```
/*st106s01.sas*/

%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces
        Season_Sold Garage_Type_2 Foundation_2 Heating_QC
        Masonry_Veneer Lot_Shape_2 Central_Air;


/*In this example, the data set ameshousing3 is divided into */
/*training and validation using the PARTITION statement, */
/*along with the SEED= option in the PROC GLMSELECT statement.*/
proc glmselect data=STAT1.ameshousing3
               plots=all
               seed=8675309;
   class &categorical / param=ref ref=first;
   model SalePrice=&categorical &interval /
                   selection=stepwise
                   (select=aic
                   choose=validate) hierarchy=single;
   partition fraction(validate=0.3333);
   title "Selecting the Best Model using Honest Assessment";
run;
```

   Here are the results.

2. Which model did PROC GLMSELECT choose?

   PROC GLMSELECT chose the model at Step 10, which has the following effects:**Intercept**, **Basement_Area**, **Gr_Liv_Area**, **Age_Sold**, **Garage_Area**, **Overall_Cond2**, **Fireplaces**, **Overall_Qual2**, **House_Style2**, **Deck_Porch_Area**, and **Heating_QC**.

3. Resubmit the PROC GLMSELECT step. Do not make any changes to it. Does it produce the same results as before?

The results are the same. Every time you run a specific PROC GLMSELECT step using the same seed value, the pseudo-random selection process is replicated and you get the same results.

4. In the PROC GLMSELECT statement, change the value of SEED= and submit the modified code. Does it produce the same results as before?

   Because you used a different seed, the results are almost certainly different from the previous results.

Hide Solution