Print

Demo: Performing Stepwise Regression Using PROC GLMSELECT

Filename: **st104d01.sas**

In this demonstration, we use the PROC GLMSELECT STEPWISE selection method to produce candidate models for predicting SalePrice.

PROC GLMSELECT DATA=*SAS-data-set* <*options*>;
  CLASS *variable(s)*;
  <*label:*>**MODEL** *dependent* = <*effects*> < / *options*>;
RUN;

1. Open program st104d01.sas.

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

/*st104d01.sas*/
ods graphics on;
proc glmselect data=STAT1.ameshousing3 plots=all;
        STEPWISE: model SalePrice = &interval / selection=stepwise details=steps select=SL slstay=0.
        title "Stepwise Model Selection for SalePrice - SL 0.05";
run;


/*Optional Code that will execute forward and backward selection
  Each with slentry and slstay = 0.05.

proc glmselect data=STAT1.ameshousing3 plots=all;
        FORWARD: model SalePrice = &interval / selection=forward details=steps select=SL slentry=0.(
        title "Forward Model Selection for SalePrice - SL 0.05";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
        BACKWARD: model SalePrice = &interval / selection=backward details=steps select=SL slstay=0.
        title "Backward Model Selection for SalePrice - SL 0.05";
run;
*/
```

The PROC GLMSELECT statement specifies the ameshousing3 data set, and the PLOTS=ALL option requests all of the available ODS plots that are generated by the provided statements. If you wanted to specify a categorical predictor variable, you would include a CLASS statement, as shown in the syntax. The MODEL statement specifies SalePrice as the response variable and all of the variables in the interval macro variable as the predictor variables. Notice the corresponding label in the MODEL statement. This label assists us when looking at the code. It's not part of the output. You'll see how useful this is in the next demonstration, when we use multiple PROC steps.

The option SELECTION=stepwise tells PROC GLMSELECT to use the stepwise selection method to select the model. If you omit this option, the procedure uses the STEPWISE method by default. We also use the DETAILS=steps to display a table and graph of the entry candidates for each step of the selection process. Next, SELECT= specifies the criterion that determines the order in which effects enter or leave at each step of the specified selection method. The default value is SELECT=SBC. We'll use SL for Significance Level as the criterion for selecting variables for the model. The default significance level for a variable to enter or stay in the model using stepwise selection is 0.15. Here, we change these significance levels by adding SLSTAY=0.05 and SLENTRY=0.05.

2. Submit the code.

3. Review the output.

The first part of the output, ModelInfo, provides basic information about the model selection process such as the specified significance levels for variables to enter and stay in the model. Next, in StepDetails, SAS displays information about the model at each step of the process. The process begins at step 0 with the empty model--that is, the model that contains the intercept only. At Step 1, Basement_Area entered the model first, which indicates that of all the one-variable models, Basement_Area was the most significant.

You can see the ANOVA table and Fit Statistics the includes summary statistics for the model at Step 1. The Parameter Estimates table provides information about the variables that are in the Step 1 model. You can display p-values in the Parameter Estimates table by including the SHOWPVALUES option in the MODEL statement.

The Candidates table and Candidates Plot list the variables that were candidates for entry into the model at this step based on their significance level. The variables are ranked, starting with the variable with the most significant p-value under the threshold of the specified entry level. Notice that the p-values of most of the variables are listed as simply less than .0001. To distinguish between these candidates, the log of the p-value for each effect is also displayed. From both the table and the graph, you see that Basement_Area is first to enter the model.

At Step 2, Above Grade Living Area entered the model. At this point, the selection method checks whether the first predictor, Basement_Area, has become non-significant, and if so, removes it. Basement_Area remained significant, so at Step 3, Age_Sold entered the model.

Let's scroll down to the Selection Summary table, Stepwise Selection Summary. Here you can see the variable that was added at each step of the process. The F values and p-values shown in this summary table are not the F and p values for the selected model. These are statistics for each individual step. Final p-values

and parameter estimates can be found in the table preceding this summary or at the conclusion of the output. No variables were removed from the model. At each step, SAS checked for any variables in the model that became non-significant, but found none. A note explains that the selection process stopped because the candidate for entry has an SLE greater than 0.05 and the candidate for removal has an SLS less than 0.05.

The Stop Details table shows that the next candidate for entry, Total_Bathrooms, did not meet the criterion for entry, and the next candidate for removal, Lot_Area, did not meet the criterion for removal.

Next, in the Coefficient Panel, we see the Coefficient Progression for SalePrice panel. PROC GLMSELECT displays a panel of two plots that show how the standardized coefficients and the criterion were used to choose the final model that evolved as the selection progressed. The green (or red) line tracks the first term that was added to the model, Basement_Area. When Basement_Area entered the model, it had a standardized coefficient of slightly more than 0.6. When the next variable entered the model, the standard coefficient for Basement_Area decreased. This graph keeps a running track of the standardized coefficients, and you can see when they start to stabilize.

The Criteria Panel lists values for a default set of fit criteria, AIC, SBC, AICC, and adjusted R-square.The star denotes the best model of the eight that were tested in this example. The model at Step 7 has the best AIC, AICC, and adjusted R-square, and the model at Step 6 has the best SBC.

The Progression of Average Squared Errors, ASE plot, shows the average squared error (ASE) evaluated for the model that was chosen at each step. Average squared error is similar to mean squared error, but it uses a different denominator in its calculation. As more effects are added to the model, the ASE decreases. There is less unexplained variation in SalePrice. When sample data is split into multiple data sets for model development and model comparison or testing, this plot shows the ASE for each of those data sets. This plot shows small differences in the unexplained variation between the models at Steps 5, 6, and 7. The stepwise process based on significance level chose the model at Step 7. But if you chose to use the model at Step 6 or even 5, you could do so with little loss of explanatory power.

The last section of output, Selected Model, provides information about the selected model, including the list of predictor variables. But remember, the p-values are likely biased because they came from automatic model selection. The bootstrap method could help calculate more appropriate standard errors, and consequently, unbiased p-values.

Stepwise selection efficiently searched through a subset of possible models using significance level as the criterion for variable selection, and suggested a model using seven of the eight potential predictors. The recommended model includes all predictors except Total Bathrooms, so these are the ones to consider when you predict the sale price of homes. Model fit plots compared the models at each step of the stepwise process, including the intercept-only model. The recommended model has the highest adjusted R-square (0.8046) and optimal values of AIC and AICC. The optimal SBC was for a six-predictor model that lacked Lot_Area.

Close