Print

Demo: Building a Predictive Model Using PROC GLMSELECT

Filename: **st106d01.sas**

In this demonstration we use the GLMSELECT procedure to build a predictive linear regression model of SalePrice from both categorical and continuous predictors. Remember that there are other SAS procedures that build predictive models. You might want to explore them outside this demonstration.

Using honest assessment, PROC GLMSELECT can build a model in two ways. If your data is already partitioned into training and validation data sets, you can simply reference both data sets in the procedure. If you start with a single data set, PROC GLMSELECT can partition the data for you. In this demo, we have a validation data set, so we won't partition our data.

PROC GLMSELECT DATA=*SAS-data-set* <**VALDATA=***validation-data-set*> <*options*>;
    CLASS *variables*;
    MODEL *target(s)* = *input(s)* < / *option(s)*>;
    STORE <OUT=> *item-store-name* < / **LABEL=***'label'*>;
RUN;

1. Open program st106d01.sas.

```
/*st106d01.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;

ods graphics;

proc glmselect data=STAT1.ameshousing3
               plots=all
               valdata=STAT1.ameshousing4;
    class &categorical / param=glm ref=first;
    model SalePrice=&categorical &interval /
               selection=backward
               select=sbc
               choose=validate;
    store out=STAT1.amesstore;
    title "Selecting the Best Model using Honest Assessment";
run;
```

In the PROC GLMSELECT statement, DATA= names the training data set, ameshousing3, and VALDATA= identifies the validation data set, ameshousing4. Ameshousing4 is a separate sample of 300 homes sold in Ames, Iowa from 2006 to 2010 that we'll use to compare model performance. PLOTS=ALL displays all ODS plots that PROC GLMSELECT produces.

The CLASS statement uses a macro variable reference to identify the categorical variables. PARAM=

specifies GLM parameterization, where one design variable, also known as a dummy or indicator variable, is produced for each level of each CLASS variable. REF=FIRST causes the first level of the CLASS variable to be treated as the redundant level, and the parameter estimates of all other levels will be compared to that reference level. The default is REF=LAST.

In the MODEL statement, SalePrice is the target, and we reference macro variables to specify the categorical and interval inputs. We use backward selection, and SBC (the Schwarz-Bayesian criterion) to determine which variables remain in the model. SBC is the default criterion for PROC GLMSELECT. CHOOSE=VALIDATE specifies that the best model will be selected based on the smallest average squared error, or ASE, for the validation data. The ASE is simply the average of the squared differences between the observed values and the predicted values using the model, that is, the sum of squared errors divided by n, the sample size.

The STORE statement creates amesstore in the stat1 library. We'll use this permanent item store to score new data in the next demonstration.

2. Submit the code.

3. Review the output.

   The first table, the Model Info table, summarizes model information including the data sets, dependent variable, selection method, and the criteria that are used to select the variables and the final model. The last row of this table refers to effect hierarchy. Maintaining model hierarchy is necessary only when you consider higher order terms in the model, such as polynomials or interactions. Enforcing effect hierarchy simply means lower order terms must be in the model if higher order terms are in the model. For example, if a quadratic term is in the model, its corresponding linear term must also be in the model.

   There are two NObs or Observation Profile tables, one for the analysis or training data, and one for the validation data. Both input data sets contain 300 observations. We see that 294 observations were used for training, but only 293 observations were used for validation. This indicates that six observations in the training data set had missing information and seven had missing values in the validation data set.

   The Class Level Information table displays the 11 categorical variables that are included in the initial model. Notice, in the Dimensions table, that the number of parameters is much larger than the number of effects. This is because categorical variables have at least one parameter per effect. In the Dimensions table, the number of effects refers to the total number of continuous variables, 8, and the total number of categorical variables, 11, which are considered for model development, plus one effect for the intercept. The number of parameters refers to the total number of continuous variables, again 8, plus the total number of levels for the categorical variables, 34, plus one parameter for the intercept. In total, 43 parameters are shown in the Dimensions table. The 34 parameters for the categorical variables correspond to the 34 dummy variables that are used to distinguish among the 34 total levels.

   Now let's look at the model selection information. In the Backward Selection Summary table, Step 0 shows that we start with 20 effects and 32 parameters. Notice that this is different from the 43 parameters in the Dimensions table. The 32 parameters here refer to the initial total degrees of freedom, that is, the non-redundant parameters. Each categorical variable has one redundant dummy variable. Therefore, the total number of variables, 43, minus the number of categorical variables, 11, equals the 32 non-redundant parameters.

   The Schwarz-Bayesian criterion is assessed on the training data. In Step 1, you can see that Season_Sold was removed first because it produced the largest reduction in the SBC. Although Season_Sold has four levels (spring, summer, winter, and fall), notice that it only removed three parameters at Step 1 because Season_Sold is associated with only three degrees of freedom. For example, if we know the home was not sold in spring, summer, or winter, then, of course, it was sold in the fall. That is, the fourth level of Season_Sold provides redundant information, so only three parameters are required.

   Moving down the SBC column, you see that variables continue to be removed as long as the value of SBC continues to decrease. In the last row, Step 8, the SBC value is followed by an asterisk, which

indicates Optimal Value of Criterion.

Remember that the SBC is also the stopping criterion. So, based on the SBC for the training data, the model at Step 8, where Lot_Shape_2 was removed, is the best model according to the backward selection process and chosen criteria.

The next column, ASE, reports the training average squared error. Notice that, on the training data set, the best ASE, where smaller is better, is the model with all parameters included in the model. However, the model is chosen based on the Validation ASE to ensure that the chosen model generalizes well to new data. Looking at the Validation ASE column, we see that the ASE continues to decrease until we get to Step 6, when Heating_QC is removed. At that point, the values start to increase. The model in Step 5, when Central_Air is removed, is marked with an asterisk, which indicates the best model based on the validation data. Therefore, the best model, according to the validation ASE, has removed Season_Sold, House_Style2, Foundation_2, Garage_Type2, and Central_Air from the total set of candidate predictors.

Let's look at the Coefficient Panel. In the top section of the Coefficient Progression for SalePrice plot you can see how the parameters changed over time as variables were removed. The lower section shows the performance of the eight models based on the validation average squared error. The vertical line references the selected model at step 5. The Validation ASE values at steps 4 and 5 are relatively close to each other. Not only does the model at step 5 correspond to a slightly better validation ASE, but it's also a more parsimonious model because it has one less predictor.

In the ASE Plot, the Progression of Average Squared Errors by Role for SalePrice plot shows the ASE progression for the training data on the bottom, and validation data on the top throughout the backward selection process. The validation ASE is identical to the previous plot. Unlike the validation ASE, the training ASE doesn't decrease. It only increases because the backward elimination criterion is used to remove variables from the model. For the training data, the ASE can decrease only if variables are added.

The effects listed in the Selected Effects table display the inputs that were chosen according to the validation ASE at step 5 of the process. The Analysis of Variance (ANOVA) table displays the degrees of freedom, sums of squares, mean squares, and the F value for the final model at step 5. Below the ANOVA table, the Fit Statistics table reports summary information for the final model as well, including adjusted R-square, AICC, and the ASE values that we saw earlier.

The Parameter Estimates table displays the estimated parameters for the final model. Remember that we used the option REF=FIRST in the CLASS statement. This means that the first level of each categorical variable is always set to zero so that we don't have an over-parameterized model. For example, Overall_Qual2 has three levels, 4, 5, and 6. Level 4 is the reference level, which makes it the redundant design variable.

Overall_Qual2 5 is a design variable that compares level 5 to the reference level (level 4). Likewise, Overall_Qual2 6 is a design variable that compares level 6 to level 4. The Estimate value for Overall_Qual2 5 is the amount of change in the response variable when all other predictor variables are held constant, and Overall_Qual2 is changed from level 4 to level 5. Again, we need only two parameters for a three-level categorical variable because we only have two degrees of freedom. If the Overall_Qual2 is not level 5 or 6, then it must be level 4.

The Parameter Estimates table also shows t values. We could request p-values using the SHOWPVALUES option in the MODEL statement to view the significance of the effects in this final model. Using backward selection to consider a variety of models fit on training data, and the validation ASE metric to choose a predictive model, we reduced the initial model with 19 total variables to our final model with only 14 variables after removing Season_Sold, House_Style_2, Foundation_2, Garage_Type_2, and Central_Air. Removing these five variables improved the models predictive performance when measured against the validation data set. This also provides a flexible and generalizable model. We can now move forward and use the chosen model to score new data.

Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression

Close