Paper 207-31

Introducing the GLMSELECT PROCEDURE for Model Selection

Robert A. Cohen, SAS Institute Inc. Cary, NC

ABSTRACT

This paper describes the GLMSELECT procedure, a new procedure in SAS/STAT software that performs model selection in the framework of general linear models. This procedure supports a variety of model selection methods, including the LASSO method of Tibshirani (1996) and the related LAR method of Efron et al. (2004). The procedure enables selection from a very large number of effects (tens of thousands) and offers extensive capabilities for customizing the selection with a wide variety of selection and stopping criteria.

INTRODUCTION

When faced with a predictive modeling problem that has many possible predictor effects—dozens, hundreds, or even thousands—a natural question is “What subset of the effects provides the best model for the data?” Statistical model selection seeks to answer this question, employing a variety of definitions of the “best” model as well as a variety of heuristic procedures for approximating the true but computationally infeasible solution. The GLMSELECT procedure implements statistical model selection in the framework of general linear models. Methods include not only extensions to GLM-type models of methods long familiar in the REG procedure (forward, backward, and stepwise) but also the newer LASSO and LAR methods of Tibshirani (1996) and Efron et al. (2004), respectively.

Note that while the model selection question seems reasonable, trying to answer it for real data can lead to problematic pitfalls, including

- Only one model is selected, and even that is not guaranteed to be the “best”; there may be other, more parsimonious or more intuitively reasonable models that may provide nearly as good or even better models, but which the particular heuristic method employed does not find.
- Model selection may be unduly affected by outliers.
- There is a “selection bias” because a parameter is more likely to be selected if it is above its expected value than if it is below its expected value.
- Standard methods of inference for the final model are invalid in the model selection context.

To some degree, these pitfalls are intrinsic, and they have even led some experts to stridently denounce model selection. However, certain features of GLMSELECT, in particular the procedure’s extensive capabilities for customizing the selection and its flexibility and power in specifying complex potential effects, can partially mitigate these problems.

The main features of the GLMSELECT procedure are as follows:

- **Model Specification**
  - offers different parameterizations for classification effects
  - supports any degree of interaction (crossed effects) and nested effects
  - supports hierarchy among effects
  - provides for internal partitioning of data into training, validation, and testing roles
- **Selection Control**
  - provides multiple effect selection methods

  - enables selection from a very large number of effects (tens of thousands)
  - offers selection of individual levels of classification effects
  - provides effect selection based on a variety of selection criteria
  - provides stopping rules based on a variety of model evaluation criteria
  - provides leave-one-out and k-fold cross validation

- **Display and Output**
  - produces graphical representation of selection process
  - produces output data sets containing predicted values and residuals
  - produces macro variables containing selected models
  - supports parallel processing of BY groups
  - supports multiple SCORE statements

MODEL SELECTION METHODS

The GLMSELECT procedure extends the familiar forward, backward, and stepwise methods as implemented in the REG procedure to GLM-type models. Quite simply, forward selection adds parameters one at a time, backward elimination deletes them, and stepwise selection switches between adding and deleting them. You can find details of these methods in the PROC GLMSELECT and PROC REG documentation. In addition to these methods, PROC GLMSELECT also supports the newer LASSO and LAR methods. In the “Customizing the Selection Process” section on page 3 you can find details of how all these methods can be customized using a variety of fit criteria that are described in the “Criteria Used in Model Selection Methods” section on page 4.

LEAST ANGLE REGRESSION (LAR)

Least angle regression was introduced by Efron et al. (2004). Not only does this algorithm provide a selection method in its own right, but with one additional modification it can be used to efficiently produce LASSO solutions. Just like the forward selection method, the LAR algorithm produces a sequence of regression models where one parameter is added at each step, terminating at the full least-squares solution when all parameters have entered the model.

The algorithm starts by centering the covariates and response and scaling the covariates so that they all have the same corrected sum of squares. Initially all coefficients are zero, as is the predicted response. The predictor that is most correlated with the current residual is determined and a step is taken in the direction of this predictor. The length of this step determines the coefficient of this predictor and is chosen so that some other predictor and the current predicted response have the same correlation with the current residual. At this point, the predicted response moves in the direction that is equiangular between these two predictors. Moving in this direction ensures that these two predictors continue to have a common correlation with the current residual. The predicted response moves in this direction until a third predictor has the same correlation with the current residual as the two predictors already in the model. A new direction is determined that is equiangular between these three predictors, and the predicted response moves in this direction until a fourth predictor joins the set having the same correlation with the current residual. This process continues until all predictors are in the model.

LASSO SELECTION (LASSO)

LASSO (Least Absolute Shrinkage and Selection Operator) selection arises from a constrained form of ordinary least squares where the sum of the absolute values of the regression coefficients is constrained to be smaller than a specified parameter. More precisely, let $X = (x_1, x_2, \ldots, x_m)$ denote the matrix of covariates and let y denote the response, where the x_is have been centered and scaled to have unit standard deviation and mean zero, and y has mean zero. Then, for a given parameter t, the LASSO regression coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ are the solution to the constrained optimization problem

$$\text{minimize}||y - X\beta||^2 \qquad \text{subject to} \quad \sum_{j=1}^{m} |\beta_j| \leq t$$

Provided that the LASSO parameter t is small enough, some of the regression coefficients will be exactly zero. Hence, you can view the LASSO as selecting a subset of the regression coefficients for each LASSO parameter. By increasing the LASSO parameter in discrete steps you obtain a sequence of regression coefficients where the nonzero coefficients at each step correspond to selected parameters.

Early implementations (Tibshirani 1996) of LASSO selection used quadratic programming techniques to solve the constrained least-squares problem for each LASSO parameter of interest. Later Osborne, Presnell, and Turlach (2000) developed a “homotopy method” that generates the LASSO solutions for all values of t. Efron et al. (2004) derived a variant of their algorithm for least angle regression that can be used to obtain a sequence of LASSO solutions from which all other LASSO solutions can be obtained by linear interpolation. This algorithm for SELECTION=LASSO is used in PROC GLMSELECT. It can be viewed as a stepwise procedure with a single addition to or deletion from the set of nonzero regression coefficients at any step.

CUSTOMIZING THE SELECTION PROCESS

All of the selection methods produce a sequence of models with effects selected in various ways. You can use the SELECT= option to customize how these effects are selected, the STOP= option to customize how to quit producing this sequence, and the CHOOSE= option to customize which model in the sequence is chosen as the final model. The criteria that you can use with these options are described in the “Criteria Used in Model Selection Methods” section on page 4.

THE SELECT= OPTION

In the traditional implementations of forward, backward, and stepwise selection, the statistic used to gauge improvement in fit when an effect is added or dropped is an F statistic that reflects that effect’s contribution to the model. Note that because effects can contribute different degrees of freedom to the model, comparisons are made using p-values corresponding to these F statistics. A well-known problem with this methodology is that these F statistics do not follow an F distribution (Draper, Guttman, and Kanemasu 1971). Hence these p-values cannot reliably be interpreted as probabilities. You can use the SELECT= option to specify an alternative statistic for gauging improvements in fit.

For example, if you specify

```
   selection=backward(select=AICC)
```

then at any step of the selection process, the effect whose removal yields the smallest “Corrected Akaike Criterion (AICC)” is the effect that gets dropped at that step.

THE STOP= OPTION

By default, the statistic used to terminate the selection process is the same statistic that is used to select the sequence of models. When p-values are used in the traditional forward, backward, and stepwise methods, selection stops when all entering effects are not significant at a prespecified “Significance Level for Entry (SLE)” and all effects in the model are significant at a prespecified “Significance Level to Stay (SLS).” In addition to the aforementioned problem that these significance levels are not reliably interpreted as probabilities, another problem with this approach is that the SLE and SLS values do not depend on the data. Thus, the same entry significance level can result in overfitting for some data and underfitting for other data. You can address this issue by using the STOP= option to specify an alternative statistic for terminating the selection process.

For example, if you specify

```
   selection=forward(select=SL stop=PRESS)
```

then effects are added to the model based on significance level but the selection process terminates when adding any effect to the model increases the predicted residual sum of squares (PRESS).

THE CHOOSE= OPTION

The CHOOSE= option enables you to specify a criterion to use for picking a model from the sequence of models obtained by the selection process. If you do not specify a CHOOSE= criterion, then the model based on the STOP= option is the selected model.

For example, if you specify

```
   selection=lasso(choose=CP)
```

then a sequence of models is obtained using the LASSO algorithm. From the models in this sequence, the one yielding the smallest value of the Mallow's $C(p)$ statistic is chosen as the final model.

CRITERIA USED IN MODEL SELECTION METHODS

PROC GLMSELECT supports a variety of fit statistics that you can specify as criteria for the CHOOSE=, SELECT=, and STOP= options in the MODEL statement. The following statistics are available:

ADJRSQ	the Adjusted R-square statistic (Darlington 1968; Judge et al. 1985)
AIC	the Akaike Information Criterion (Darlington 1968; Judge et al. 1985)
AICC	the Corrected Akaike Information Criterion (Hurvich and Tsai 1991)
BIC	the Sawa Bayesian Information Criterion (Sawa 1978; Judge et al. 1985)
CV	the $k-$fold cross validation predicted residual sum of squares
CP	the Mallow C(p) statistic (Mallows 1973; Hocking 1976)
PRESS	the predicted residual sum of squares statistic (leave-one-out cross validation)
SBC	the Schwarz Bayesian Information Criterion (Schwarz 1978; Judge et al. 1985)
SL	the significance level of the F statistic used to assess an effect's contribution to the fit when it is added or dropped
VALIDATEASE	the average square error over the validation data

You can find further discussion and formula for these criteria in the PROC GLMSELECT documentation.

EXAMPLE

The following example uses simulated data to illustrate how you can use PROC GLMSELECT in model development and exploit its facilities to avoid some of the pitfalls of traditional implementations of variable selection methods. See the section "Model Selection Issues" in the PROC GLMSELECT documentation for a discussion of issues that arise with variable selection.

In this example, you reserve one-third of your data as a test data set and your goal is to develop a model using the remaining two-thirds, the training data. You assess your model on its predictive performance on the test data. To facilitate this assessment, PROC GLMSELECT computes the average square error (ASE) separately for the observations used for training and testing for the models at each step of the selection process. The ASE for observations in a given role at any selection step is the residual sum of squares of the observations in that role at that step divided by the number of observations in that role.

The following code produces the simulated analysis and test data sets. Note that in all the following analyses, the test data set specified using the TESTDATA= option in the PROC GLMSELECT statement plays no role in the selection process, and the same models would have been obtained if this data were not provided.

```
 data analysisData testData;
    drop i j;
    array x{20} x1-x20;

    do i=1 to 5000;

      /* Continuous predictors */
      do j=1 to 20;
         x{j} = ranuni(1);
      end;

      /* Classification variables */
      c1 = int(1.5+ranuni(1)*7);
      c2 = 1 + mod(i,3);
      c3 = int(ranuni(1)*15);

      yTrue = 2 + 5*x17 - 8*x5 + 7*x9*c2 - 7*x1*x2 + 6*(c1=2) + 5*(c1=5);
      y     = yTrue + 6*rannor(1);

      if ranuni(1) < 2/3 then output analysisData;
                         else output testData;
   end;
run;
```

This example is constructed so that the dependent variable depends linearly on both continuous variables (x1,x2,x5,x9,x17) and classification variables (c1,c2) as well as some two-way interactions of these effects.

This example is organized into seven analyses:

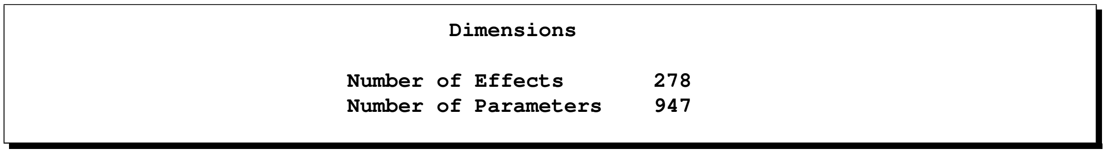
1. Forward with STOP=NONE (Full Least Squares)
2. Stepwise with SELECT=SL (Traditional Stepwise)
3. Stepwise with SELECT=SL and CHOOSE=PRESS
4. Stepwise with SELECT=SBC (Default method)
5. Stepwise with SELECT=SBC and a Split Classification Variable
6. Stepwise with SELECT=SBC STOP=VALIDATE and a Split Classification Variable
7. Stepwise with SELECT=SBC STOP=CV and a Split Classification Variable

ANALYSIS 1: FULL LEAST-SQUARES MODEL

Simply fitting a least-squares model that includes all main effects and two-way interactions produces a model that overfits your data and generalizes very poorly. The following code fits such a model using forward selection to determine the order in which effects enter the model. Because the STOP=NONE option is specified, the selection proceeds until all the specified effects are in the model. Note that if you are interested in only a full least-squares model, then it is much more efficient to specify SELECTION=NONE in the MODEL statement. However, in this example the aim is to add effects in roughly increasing order of explanatory power.

```
proc glmselect data=analysisData testdata=testData plots=asePlot;
   class c1 c2 c3;

   model y =  c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
             |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2
           / selection=forward(stop=none);
run;
```

Selecting a model from all main effects and their two-way or higher interactions often leads to a selection from a very large number of effects. Furthermore, when these effects include classification variables with several levels, then the number of parameters available for selection is even larger. The “Dimensions” table in Figure 1 shows the number of effects and the number of parameters considered in this example.



```
                    Dimensions

         Number of Effects         278
         Number of Parameters      947
```

Figure 1. Selection Dimensions

PROC GLMSELECT supports a variety of diagnostic plots that you can use to monitor the selection process. You enable these graphical displays by specifying the ODS GRAPHICS statement and you request these plots using the PLOTS=option in the PROC GLMSELECT statement. Figure 2 shows the ASE plot requested with the PLOTS=ASEPlot option. This plot tracks ASE by role as the selection process proceeds and clearly demonstrates the danger in overfitting the training data. As more insignificant effects are added to the model, the growth in test set ASE shows how the predictions produced by the resulting models worsen.

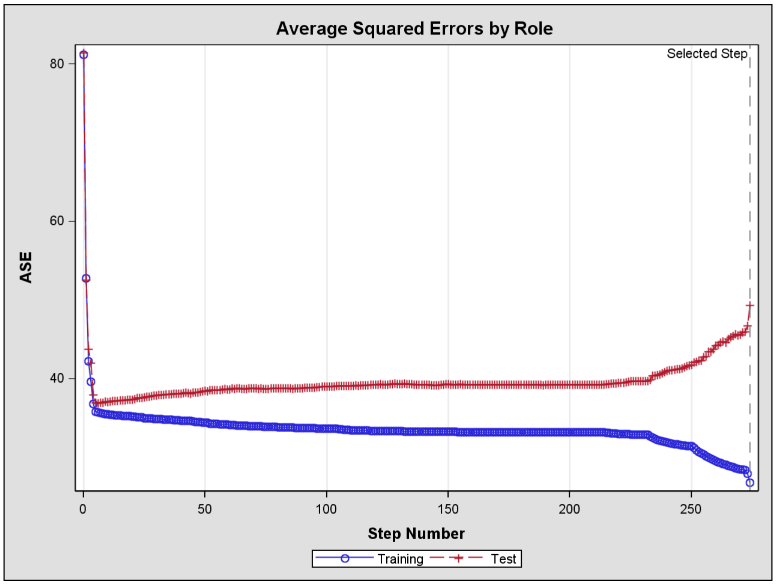


Figure 2. Average Square Errors on Training and Test Data

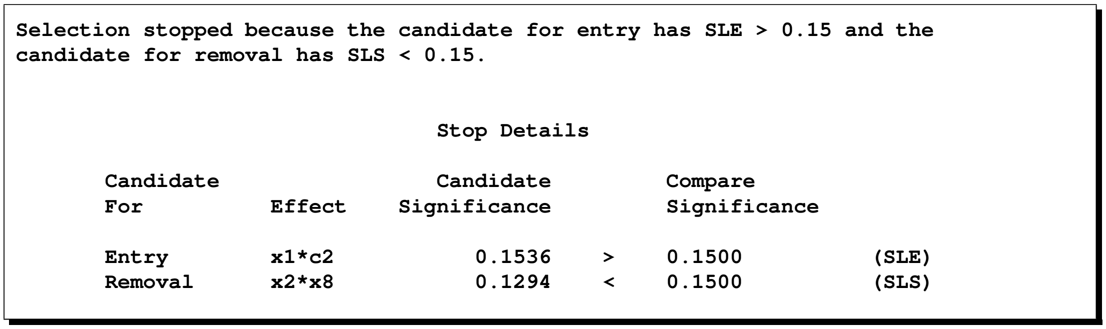
It is clear that using a model with all main effects and two-way interactions is inappropriate. Ideally, you should use a priori knowledge to determine what main effects and interactions to allow but in some cases this information might not be available. In these situations, variable selection can prove useful in finding a parsimonious model with good predictive performance.

ANALYSIS 2: TRADITIONAL STEPWISE SELECTION

The following code uses a traditional implementation of stepwise selection to obtain a model, where effects in the model that are not significant at the stay significance level (SLS) are candidates for removal, and effects not yet in the model whose addition is significant at the entry significance level (SLE) are candidates for addition to the model. You request this by specifying the SELECTION=STEPWISE(select=SL) option in the MODEL statement.

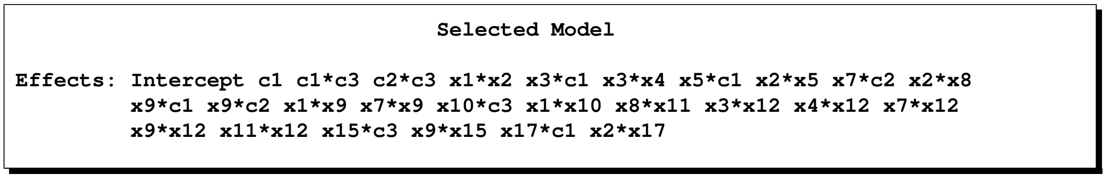
```
proc glmselect data=analysisData testdata=testData
              plots=(asePlot CoefficientPanel(unpack) Criteria);

   class c1 c2 c3;

   model y =  c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
                |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2
           / selection = stepwise(select=SL) stats=all;
run;
```

The “Stop Reason” and “Stop Details” tables in Figure 3 provide details of why the selection process terminated, and the effects in the selected model are displayed in Figure 4.



```
Selection stopped because the candidate for entry has SLE > 0.15 and the
candidate for removal has SLS < 0.15.


                            Stop Details

      Candidate                   Candidate          Compare
      For          Effect      Significance          Significance

      Entry        x1*c2             0.1536     >    0.1500           (SLE)
      Removal      x2*x8             0.1294     <    0.1500           (SLS)
```

Figure 3. Stop Details



```
                            Selected Model

Effects: Intercept c1 c1*c3 c2*c3 x1*x2 x3*c1 x3*x4 x5*c1 x2*x5 x7*c2 x2*x8
         x9*c1 x9*c2 x1*x9 x7*x9 x10*c3 x1*x10 x8*x11 x3*x12 x4*x12 x7*x12
         x9*x12 x11*x12 x15*c3 x9*x15 x17*c1 x2*x17
```

Figure 4. Selected Effects

One problem with this methodology is that the default SLE and SLS values of 0.15 are not appropriate for this data, nor as pointed out earlier can they reliably be interpreted as probabilities. The sequence of entry p-values at each step are plotted in Figure 5. Note that this sequence is not monotone increasing, and stopping when all candidate entering effects are not significant at the prespecified SLE value does not guarantee that, if the selection proceeded, effects would continue to be deemed not significant.

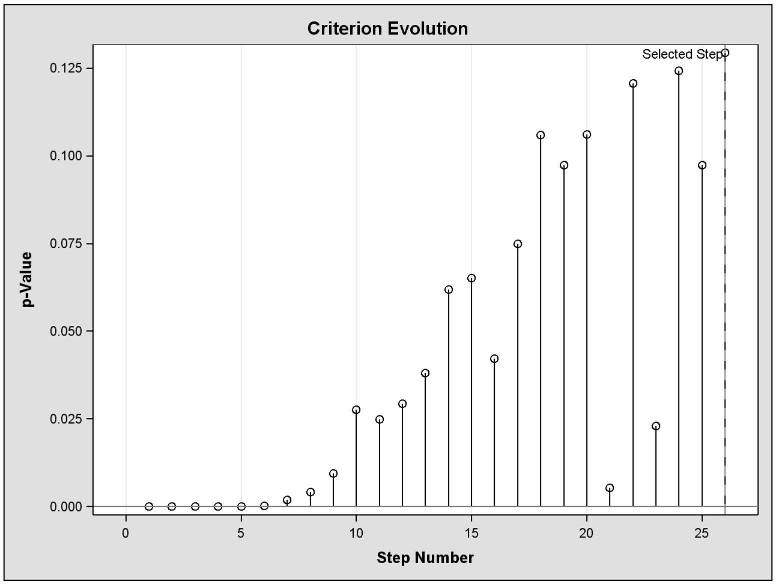


Figure 5. Entering Effect p-values

The ASE plot in Figure 6 confirms that with the default SLE and SLS values of 0.15, stepwise selection produces a model that overfits the training data.

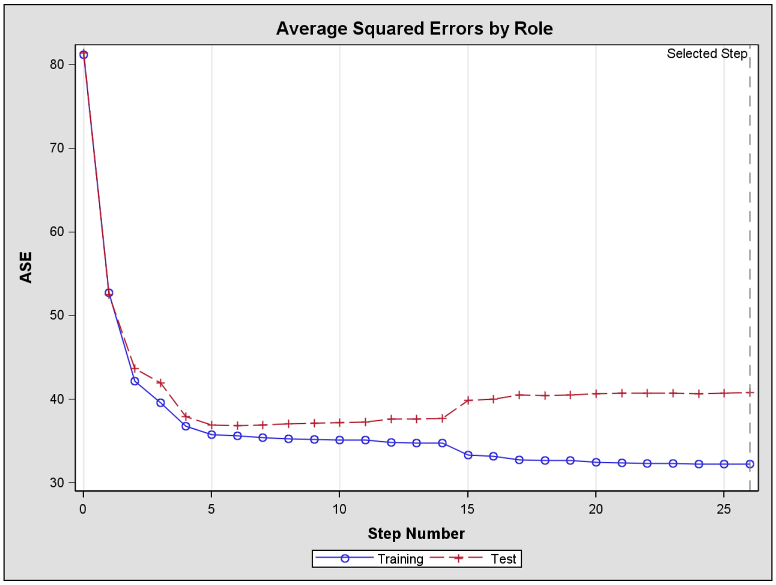


Figure 6. Average Square Errors on Training and Test Data

An alternative strategy to using significance levels to determine when to stop the selection process is to base this decision on data-driven criteria. The PLOTS=(criteria) suboption of the PLOTS=option requests the plot in Figure 7 showing the fit criteria supported by PROC GLMSELECT. With the exception of adjusted R-Square, these criteria suggest stopping the selection sooner than the actual selected step. Note that the STATS=ALL option in the MODEL statement requests that all these criteria be computed. Computing the PRESS, CP, and BIC criteria requires significant extra computation, and so these statistics are computed only if they are used in the selection process or are explicitly requested using the STATS= option.

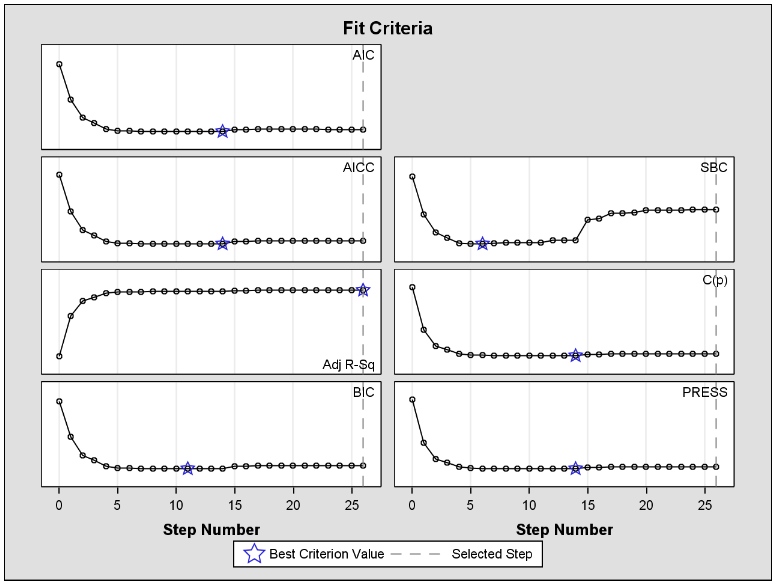


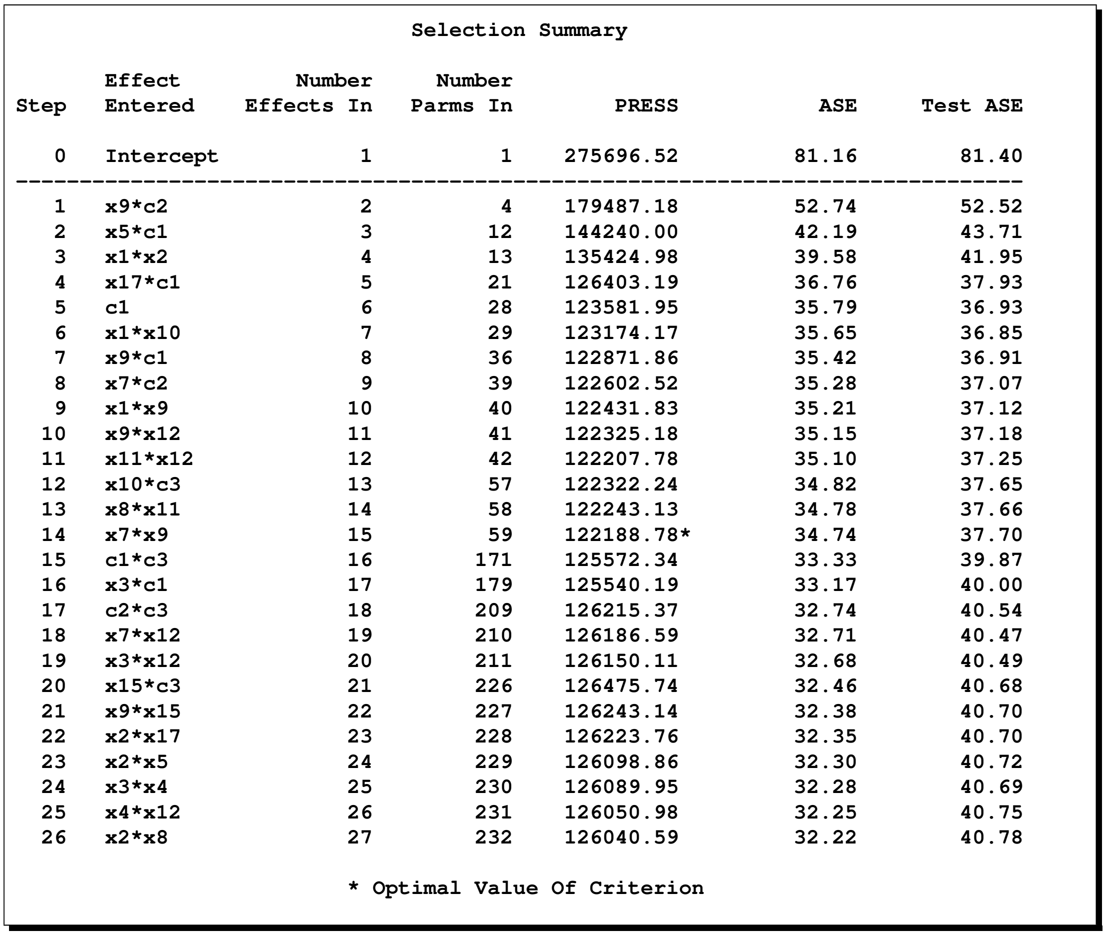
Figure 7. Fit Criteria Evolution

ANALYSIS 3: TRADITIONAL STEPWISE SELECTION WITH CHOOSE=PRESS

You can use any of the criteria shown in Figure 7 to terminate the selection process or to choose among the sequence of the models examined. The CHOOSE=PRESS suboption of the SELECTION= option in the following code requests that, among the models obtained at each step of the selection process, the final selected model is the model with the smallest predicted residual sum of squares (PRESS). Note that the CHOOSE= option does not alter the sequence of models nor when the selection terminates.

```
proc glmselect data=analysisData testdata=testData
              plots=(asePlot Criteria);

   class c1 c2 c3;

   model y =  c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
                |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2
           / selection=stepwise(select=SL choose=press);
run;
```

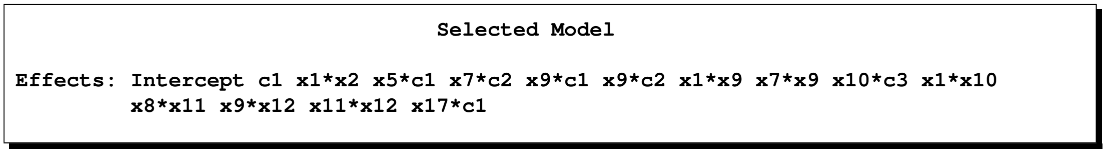
You can see how the PRESS criterion changes as the selection proceeds in the “Selection Summary” table in Figure 8.



```
                                    Selection Summary

        Effect          Number      Number
Step    Entered      Effects In    Parms In        PRESS            ASE      Test ASE

   0    Intercept             1           1    275696.52          81.16         81.40
-------------------------------------------------------------------------------------
   1    x9*c2                 2           4    179487.18          52.74         52.52
   2    x5*c1                 3          12    144240.00          42.19         43.71
   3    x1*x2                 4          13    135424.98          39.58         41.95
   4    x17*c1                5          21    126403.19          36.76         37.93
   5    c1                    6          28    123581.95          35.79         36.93
   6    x1*x10                7          29    123174.17          35.65         36.85
   7    x9*c1                 8          36    122871.86          35.42         36.91
   8    x7*c2                 9          39    122602.52          35.28         37.07
   9    x1*x9                10          40    122431.83          35.21         37.12
  10    x9*x12               11          41    122325.18          35.15         37.18
  11    x11*x12              12          42    122207.78          35.10         37.25
  12    x10*c3               13          57    122322.24          34.82         37.65
  13    x8*x11               14          58    122243.13          34.78         37.66
  14    x7*x9                15          59    122188.78*         34.74         37.70
  15    c1*c3                16         171    125572.34          33.33         39.87
  16    x3*c1                17         179    125540.19          33.17         40.00
  17    c2*c3                18         209    126215.37          32.74         40.54
  18    x7*x12               19         210    126186.59          32.71         40.47
  19    x3*x12               20         211    126150.11          32.68         40.49
  20    x15*c3               21         226    126475.74          32.46         40.68
  21    x9*x15               22         227    126243.14          32.38         40.70
  22    x2*x17               23         228    126223.76          32.35         40.70
  23    x2*x5                24         229    126098.86          32.30         40.72
  24    x3*x4                25         230    126089.95          32.28         40.69
  25    x4*x12               26         231    126050.98          32.25         40.75
  26    x2*x8                27         232    126040.59          32.22         40.78

                                * Optimal Value Of Criterion
```

Figure 8. Selection Summary

The PRESS statistic achieves a global minimum at Step 14, and the model at this step is selected. Note that the sequence of PRESS values has more than one local minimum, with the first local minimum occurring at Step 11. If you change the CHOOSE=PRESS option to STOP=PRESS, then the selection stops at Step 11 and the model at Step 11 is selected.



```
                                      Selected Model

Effects: Intercept c1 x1*x2 x5*c1 x7*c2 x9*c1 x9*c2 x1*x9 x7*x9 x10*c3 x1*x10
         x8*x11 x9*x12 x11*x12 x17*c1
```

Figure 9. Selected Model

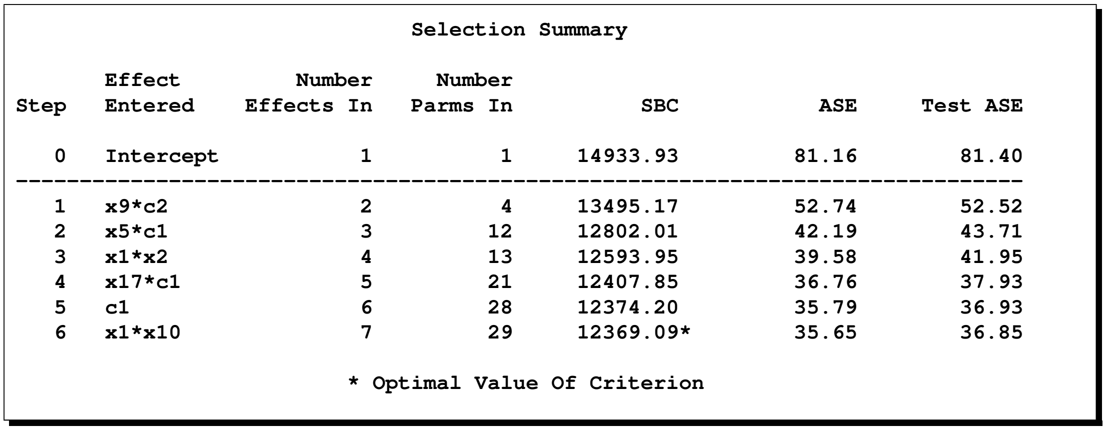
Figure 9 shows the effects in the selected model at Step 14. The ASE plot in Figure 6 shows that, by choosing this model, you limit the overfitting of the training data that occurs as selection proceeds beyond this step.

ANALYSIS 4: DEFAULT STEPWISE SELECTION (SELECT=SBC)

In the previous two analyses, independent of the STOP= or CHOOSE= specifications, the SELECT=SL option directs the traditional approach where the sequence of additions and deletions is determined by significance levels with their aforementioned problems. An alternative approach is to determine the sequence of effects using an information or out-of-sample validation criterion. By default, PROC GLMSELECT uses stepwise selection based on the Schwarz Bayesian Information Criterion (SBC). The following code uses this default with the ORDERSELECT option specifying that effects be displayed in the order in which they first entered the model.

```
proc glmselect data=analysisData testdata=testData;

   class c1 c2 c3;

   model y =  c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
                |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2
           / orderSelect;
run;
```

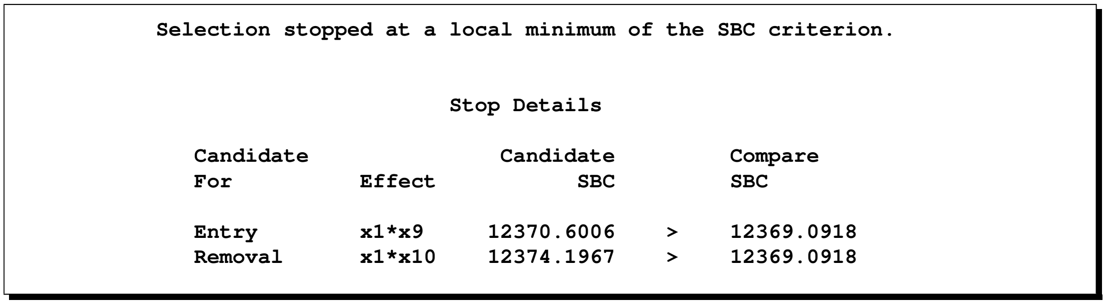
By comparing the “Selection Summary” table in Figure 10 with the selection summary table in Figure 8 you can see that in this case, using SBC as the selection criterion initially yields the same sequence of models as when the selection was based on significance levels.



```
                              Selection Summary

      Effect         Number     Number
Step  Entered      Effects In   Parms In        SBC          ASE      Test ASE

   0  Intercept             1          1    14933.93        81.16        81.40
-------------------------------------------------------------------------------
   1  x9*c2                 2          4    13495.17        52.74        52.52
   2  x5*c1                 3         12    12802.01        42.19        43.71
   3  x1*x2                 4         13    12593.95        39.58        41.95
   4  x17*c1                5         21    12407.85        36.76        37.93
   5  c1                    6         28    12374.20        35.79        36.93
   6  x1*x10                7         29    12369.09*       35.65        36.85

                         * Optimal Value Of Criterion
```

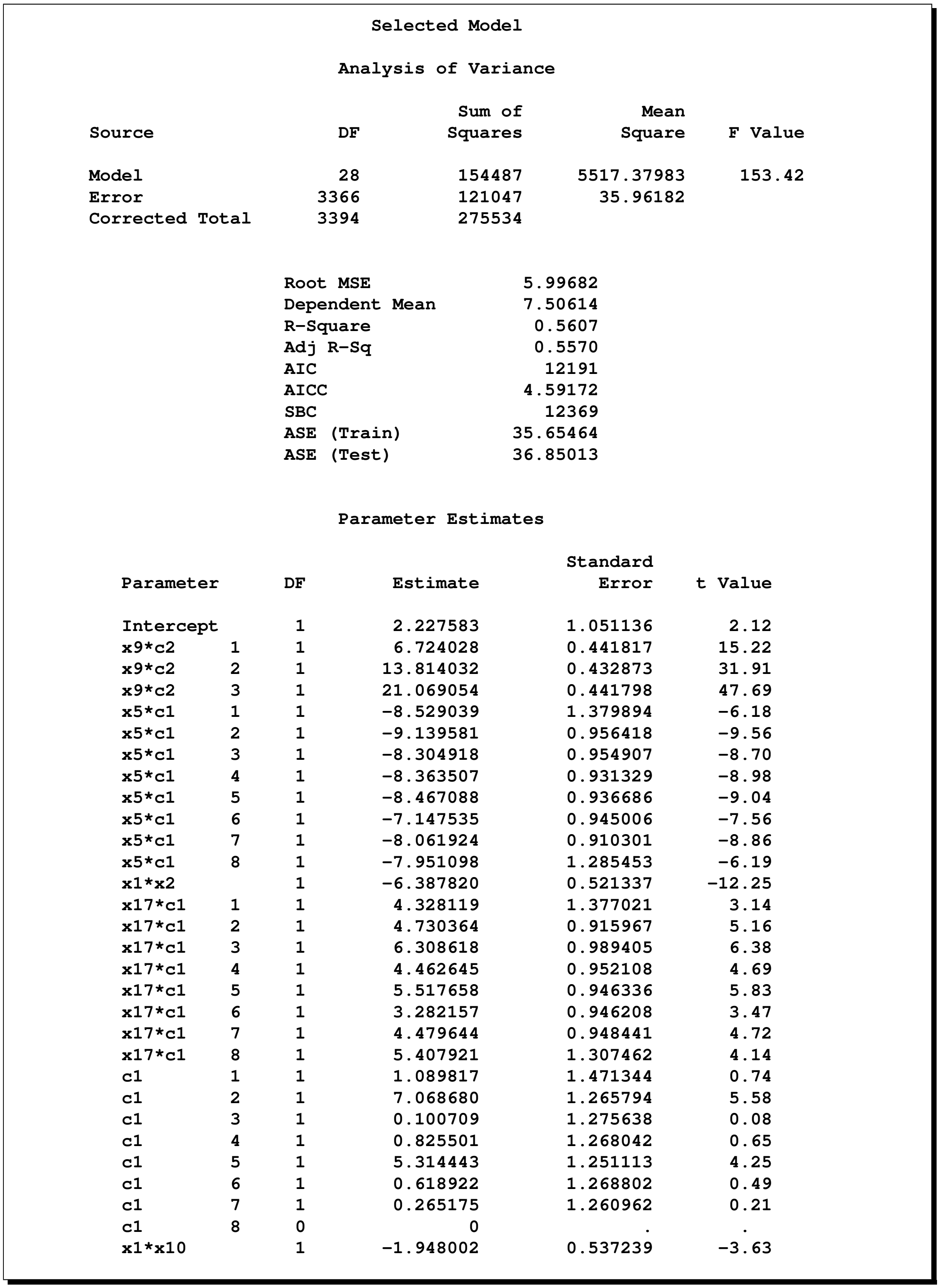
Figure 10. Selection Summary

Since no STOP= criterion is specified, the selection stepwise selection terminates when adding or dropping any effect increases the SBC statistic, as shown in Figure 11.



```
            Selection stopped at a local minimum of the SBC criterion.


                                Stop Details

               Candidate                Candidate          Compare
               For          Effect            SBC          SBC

               Entry        x1*x9      12370.6006     >    12369.0918
               Removal      x1*x10     12374.1967     >    12369.0918
```

Figure 11. Stop Details



```
                              Selected Model

                           Analysis of Variance

                                      Sum of           Mean
        Source                DF     Squares          Square    F Value

        Model                 28      154487      5517.37983     153.42
        Error               3366      121047        35.96182
        Corrected Total     3394      275534


                      Root MSE               5.99682
                      Dependent Mean         7.50614
                      R-Square                0.5607
                      Adj R-Sq                0.5570
                      AIC                      12191
                      AICC                   4.59172
                      SBC                      12369
                      ASE (Train)           35.65464
                      ASE (Test)            36.85013


                           Parameter Estimates

                                                  Standard
           Parameter       DF       Estimate         Error    t Value

           Intercept        1       2.227583      1.051136       2.12
           x9*c2      1     1       6.724028      0.441817      15.22
           x9*c2      2     1      13.814032      0.432873      31.91
           x9*c2      3     1      21.069054      0.441798      47.69
           x5*c1      1     1      -8.529039      1.379894      -6.18
           x5*c1      2     1      -9.139581      0.956418      -9.56
           x5*c1      3     1      -8.304918      0.954907      -8.70
           x5*c1      4     1      -8.363507      0.931329      -8.98
           x5*c1      5     1      -8.467088      0.936686      -9.04
           x5*c1      6     1      -7.147535      0.945006      -7.56
           x5*c1      7     1      -8.061924      0.910301      -8.86
           x5*c1      8     1      -7.951098      1.285453      -6.19
           x1*x2            1      -6.387820      0.521337     -12.25
           x17*c1     1     1       4.328119      1.377021       3.14
           x17*c1     2     1       4.730364      0.915967       5.16
           x17*c1     3     1       6.308618      0.989405       6.38
           x17*c1     4     1       4.462645      0.952108       4.69
           x17*c1     5     1       5.517658      0.946336       5.83
           x17*c1     6     1       3.282157      0.946208       3.47
           x17*c1     7     1       4.479644      0.948441       4.72
           x17*c1     8     1       5.407921      1.307462       4.14
           c1         1     1       1.089817      1.471344       0.74
           c1         2     1       7.068680      1.265794       5.58
           c1         3     1       0.100709      1.275638       0.08
           c1         4     1       0.825501      1.268042       0.65
           c1         5     1       5.314443      1.251113       4.25
           c1         6     1       0.618922      1.268802       0.49
           c1         7     1       0.265175      1.260962       0.21
           c1         8     0              0             .          .
           x1*x10           1      -1.948002      0.537239      -3.63
```

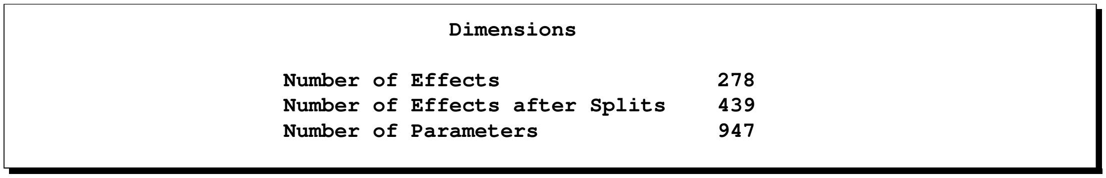
Figure 12. Details of Selected Model

Figure 12 provides details of the selected model. By default, all the parameters associated with a classification effect enter or leave the model as a unit. However, in some cases the dependent variable may depend strongly on only a subset of the levels of a classification variable. By examining the estimates and t-values of the parameters corresponding to the classification effect c1, you observe that only levels “2” and “5” of this effect contribute appreciably to the model. This suggests that a more parsimonious model with similar or better predictive power might be obtained if parameters corresponding to the levels of c1 are allowed to enter or leave the model independently. The following analysis implements this strategy.

ANALYSIS 5: DEFAULT STEPWISE SELECTION WITH A SPLIT CLASSIFICATION VARIABLE

```
proc glmselect data=analysisData testdata=testData;

   class c1(split) c2 c3;

   model y =  c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
                |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2
           / orderSelect;
run;
```

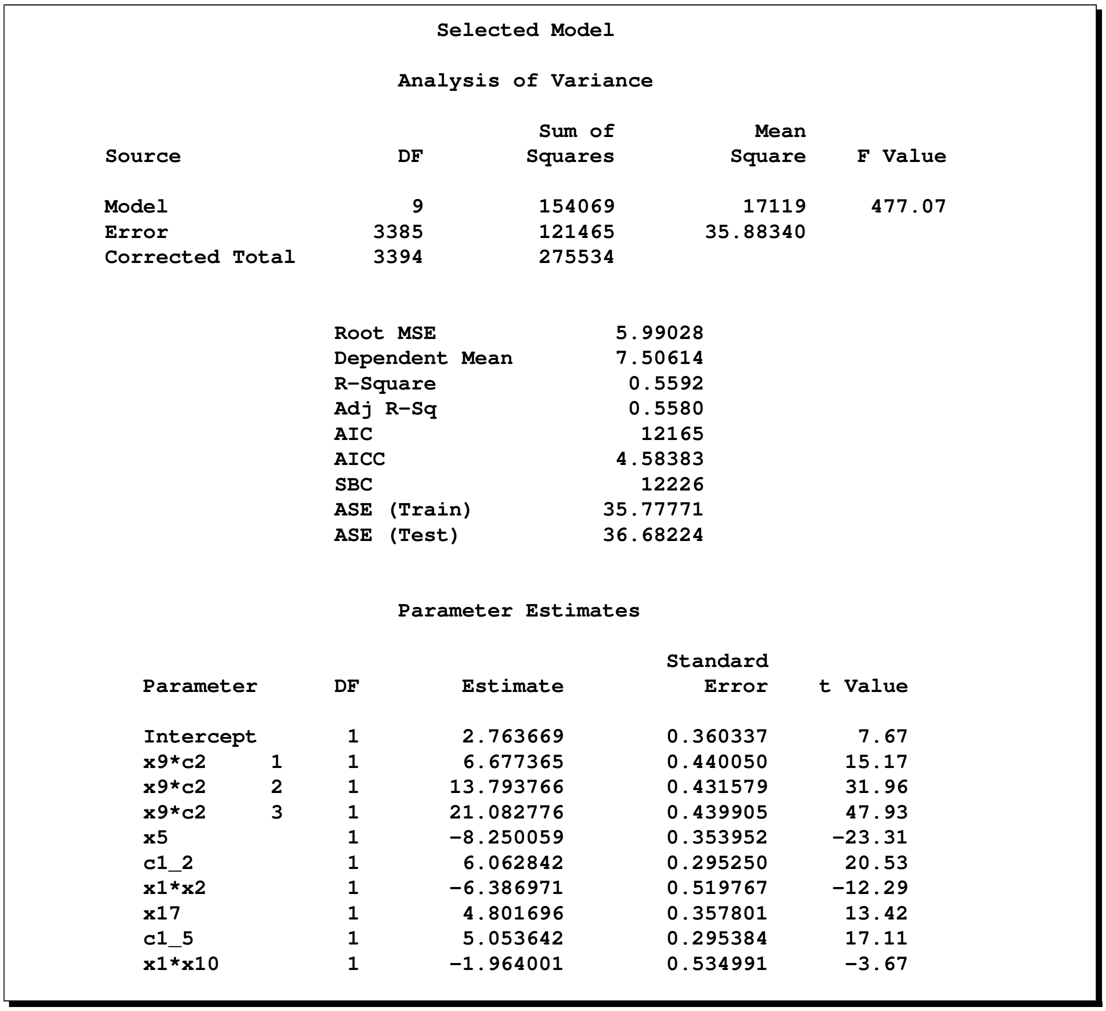
The “Dimensions” table in Figure 13 shows that, while the model statement specifies 278 effects, after splitting the parameters corresponding to the levels of c1, there are 439 split effects that are considered for entry or removal at each step of the selection process. Note that the total number of parameters considered is not affected by the split option.



```
                              Dimensions

                   Number of Effects                  278
                   Number of Effects after Splits     439
                   Number of Parameters               947
```

Figure 13. Problem Dimensions

Figure 14 provides details of the selected model. Observe that, with the split option applied to the classification effect c1, the selected models contain only two parameters corresponding to the eight levels of c1. By comparing the ANOVA and fit statistics of the selected model in Figure 14 with the corresponding ANOVA and fit statistics in Figure 12, you see that you obtain a model with fewer degrees of freedom (9 versus 28) but with improved predictive performance as measured by the average square errors on the test data (36.68 versus 36.85). Furthermore, you see that the selected model uses all the effects in the generated true model and includes only one additional interaction, x1*x10, that is not present in this underlying true model.



```
                            Selected Model

                         Analysis of Variance

                                      Sum of           Mean
Source                    DF         Squares         Square    F Value

Model                      9          154069          17119     477.07
Error                   3385          121465       35.88340
Corrected Total         3394          275534


                  Root MSE              5.99028
                  Dependent Mean        7.50614
                  R-Square               0.5592
                  Adj R-Sq               0.5580
                  AIC                     12165
                  AICC                  4.58383
                  SBC                     12226
                  ASE (Train)          35.77771
                  ASE (Test)           36.68224


                         Parameter Estimates

                                                Standard
   Parameter       DF        Estimate              Error    t Value

   Intercept        1        2.763669           0.360337       7.67
   x9*c2      1     1        6.677365           0.440050      15.17
   x9*c2      2     1       13.793766           0.431579      31.96
   x9*c2      3     1       21.082776           0.439905      47.93
   x5               1       -8.250059           0.353952     -23.31
   c1_2             1        6.062842           0.295250      20.53
   x1*x2            1       -6.386971           0.519767     -12.29
   x17              1        4.801696           0.357801      13.42
   c1_5             1        5.053642           0.295384      17.11
   x1*x10           1       -1.964001           0.534991      -3.67
```

Figure 14. Details of Selected Model

ANALYSIS 6: STEPWISE SELECTION WITH INTERNALLY PARTITIONED DATA AND STOP=VALIDATE

In the preceding analysis, the stepwise selection stopped when adding or dropping an effect increases the SBC statistic. An alternative to stopping the selection based on an information criterion is to terminate the selection using an out-of-sample-based prediction statistic. The following code uses a PARTITION statement to randomly reserve one-quarter of the observations in the analysis data set for model validation and the rest for model training.

```
proc glmselect data=analysisData testdata=testData seed=1
               plot=asePlot;

   partition fraction(validate=0.25);

   class c1(split) c2 c3;

   model y =  c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
                |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2
           / selection=stepwise(stop=validate);
run;
```

You can find details about the number of observations used for each role in the “Number of Observations” tables shown in Figure 15. You can see that of the 3,395 observations in the analysis data set, 2,551 (75.1%) observations were used for model training and 844 (24.9%) for model validation. The observations reserved for testing and provided using the TESTDATA= option are not affected by the PARTITION statement in the preceding code. However, you can use a PARTITION statement to subdivide input data for training, validation, and testing roles.

```
                      Observation Profile for Analysis Data

              Number of Observations Read                        3395
              Number of Observations Used                        3395
              Number of Observations Used for Training           2551
              Number of Observations Used for Validation          844


                        Observation Profile for Test Data

                     Number of Observations Read        1605
                     Number of Observations Used        1605
```

Figure 15. Number of Observations

The STOP=VALIDATE suboption of the SELECTION=STEPWISE option specifies that the selection process terminates when adding or dropping any effect increases the average square error on the validation data.

Figure 16 shows the progression of the average square errors separately for the training, validation, and test data. You observe the desirable behavior where the average square errors on the training, validation, and test data all decrease monotonically with the selection terminating at the step beyond which the test and validation errors would begin to grow.

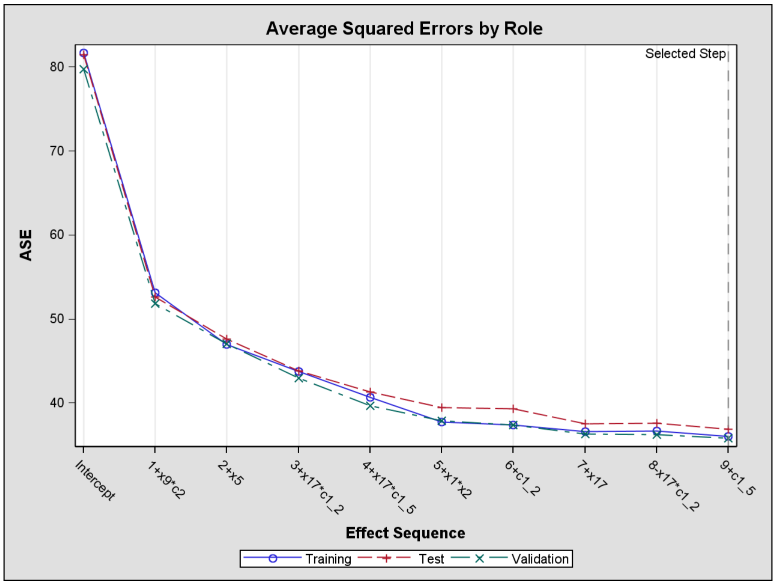


Figure 16. ASE Plot by Role

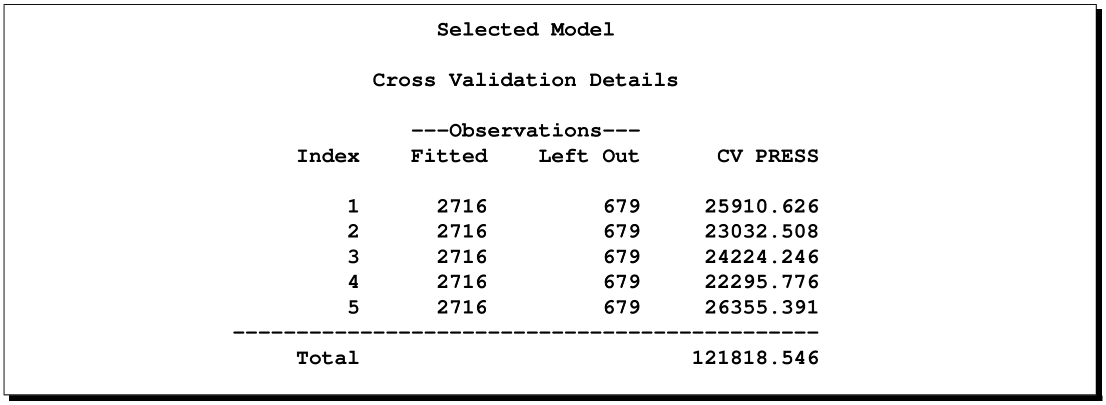
ANALYSIS 7: STEPWISE SELECTION WITH STOPPING BASED ON 5-FOLD CROSS VALIDATION

Cross validation is often used to assess the predictive performance of a model, especially when you do not have enough observations for test set validation. The following code provides an example where cross validation is used to terminate the selection.

```
proc glmselect data=analysisData testdata=testData
               plot=CoefficientPanel;

   class c1(split) c2 c3;

   model y =  c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
                |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2
           / selection=stepwise(stop=CV drop=competitive)
             cvMethod=split(5) cvDetails=all;

run;
```

The DROP=COMPETITIVE suboption of the SELECTION=STEPWISE option specifies that addition and deletion of effects are treated competitively. The selection criterion, namely SBC, is evaluated for all models obtained by deleting an effect from the current model or by adding an effect to this model. The action that most improves the SBC criterion is the action taken. This differs from the usual implementation of stepwise selection where all effects whose removal improves the selection criterion are dropped before any addition to the model is considered.

The CVMETHOD=SPLIT(5) option in the MODEL statement requests 5-fold cross validation where the training data is partitioned into five subsets consisting of observations {1,6,11,. . .}, {2,7,12,. . .}, and so on. At each step of the selection, one of these parts is held out for validation and the currently selected model is fit on the remaining four parts. This fitted model is used to compute the predicted residual sum of squares on the omitted part, and this process is repeated for each of five parts. The sum of the five predicted residual sum of squares so obtained is the estimate of the prediction error that is denoted by CVPRESS. The STOP=CV option specifies that this CVPRESS statistic is used as the stopping criterion. Figure 17 provides a breakdown of the CVPRESS statistic by fold at the final step of the selection process.



```
                         Selected Model

                    Cross Validation Details

                       ---Observations---
             Index     Fitted     Left Out        CV PRESS

                 1       2716          679       25910.626
                 2       2716          679       23032.508
                 3       2716          679       24224.246
                 4       2716          679       22295.776
                 5       2716          679       26355.391
        ----------------------------------------------
             Total                              121818.546
```

Figure 17. Cross Validation Details

The “Coefficient Panel” in Figure 18 enables you to visualize the selection process. In this plot, standardized coefficients of all the effects selected at some step of the stepwise method are plotted as a function of the step number. This enables you to assess the relative importance of the effects selected at any step of the selection process as well as providing information as to when effects entered the model. The lower plot in the panel shows how the criterion, CVPRESS, used to terminate the selection process changes as effects enter or leave the model.

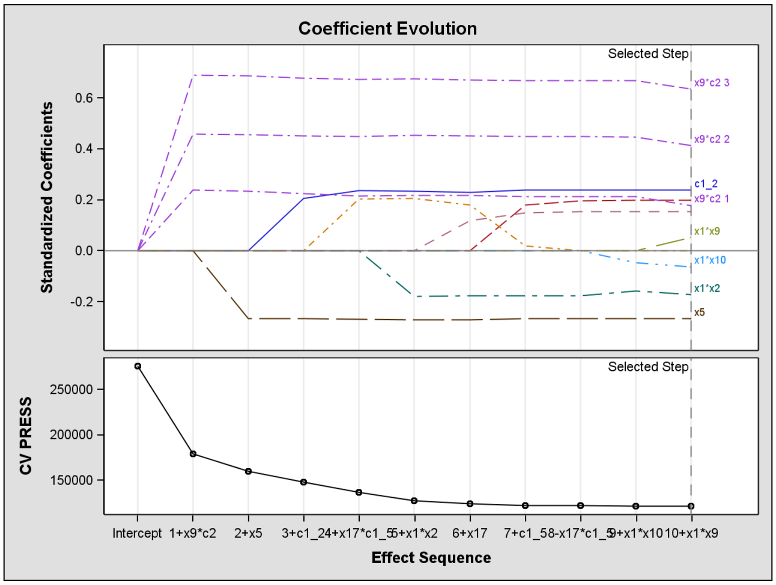


Figure 18. Coefficient Evolution

PREDICTIVE PERFORMANCE COMPARISON

The following table summarizes the performance of the final model selected by each of the seven example analyses presented. The “Effects” and “Parms” columns report the number of effects and parameters excluding the intercept. Note that for analyses 5–7 where the effect c1 is split, effects corresponding to levels of c1 are counted as a single effect. There are three columns comparing the effects in the selected model with the five effects in the underlying true model, namely x5, x17, x1*x2, c1, and x9*c2. The column “Exact” reports the number of these five effects that appear in the selected model. The column “Partial” reports the number of effects in the selected model that contain an effect in the true model but are not an exact match. The column “None” reports the number of effects in the selected model that do not contain any effect in the true model. The last row in the table gives the results for the underlying true model.

			Containment of Effects in True			ASE	
Analysis	Effects	Parms	Exact	Partial	None	Train	Test
1	274	834	5	149	120	26.73	49.28
2	26	231	3	14	9	32.22	40.78
3	14	58	3	8	3	34.74	37.70
4	5	28	2	3	0	35.65	36.85
5	6	9	5	1	0	35.77	36.68
6	6	9	5	1	0	35.99	36.88
7	7	10	5	2	0	35.71	36.72
True	5	8	5	0	0	35.96	36.73

You can see that in terms of predictive performance as measured by the average square error on the test data, the models produced by analyses 4–7 all perform comparably with each other as well as with the underlying true model. Furthermore, because the response depends only on two of the eight levels of c1, enabling the parameters corresponding to the levels of c1 to be selected independently yields the more parsimonious models of analyses 5–7. Finally, in this simulated case where the underlying true model is known, you can see that analyses 5–7 succeed in capturing all the effects in the true model without including any effect that does not contain a variable found in the true model.

OBTAINING THE GLMSELECT PROCEDURE

PROC GLMSELECT is currently an add-on to the SAS/STAT product in SAS 9.1 on the (32-bit) Windows platform. The procedure does not ship with SAS 9.1. It is downloadable from Software Downloads at support.sas.com. Documentation is also available at this site. See sas.com/statistics for up-to-date information on SAS/STAT software.

REFERENCES

Darlington, R.B. (1968), "Multiple Regression in Psychological Research and Practice," *Psychological Bulletin*, 69, 161–182.

Draper, N.R., Guttman, I., and Kanemasu, H. (1971), "The Distribution of Certain Regression Statistics," *Biometrika*, 58, 295–298.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression (with discussion)," *Annals of Statistics*, 32, 407–499.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning,* New York: Springer-Verlag, Inc.

Hocking, R.R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1–50.

Hurvich, C.M. and Tsai, C-L. (1991), "Bias of the Corrected AIC Criterion for Underfitted Regression and Time Series Models," *Biometrika*, 78, 499–509.

Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H., and Lee, T.C. (1985), *The Theory and Practice of Econometrics,* Second Edition, New York: John Wiley & Sons, Inc.

Mallows, C.L. (1967), "Choosing a Subset Regression," unpublished report, Bell Telephone Laboratories.

Mallows, C.L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

Miller, A. (2002), *Subset Selection in Regression, Second Edition*, Chapman & Hall/CRC.

Osborne, M., Presnell, B., and Turlach, B. (2000), "A New Approach to Variable Selection in Least Squares Problems," *IMA Journal of Numerical Analysis*, 20, 389–404.

Sawa, T. (1978), "Information Criteria for Discriminating Among Alternative Regression Models," *Econometrica*, 46, 1273–1282.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society Series B*, 58, 267–288.

CONTACT INFORMATION Robert Cohen, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513 Email: robert.cohen@sas.com