GLMSELECT for Model Selection

Sylvain Tremblay
SAS Canada – Education
Agenda

- Overview of GLMSELECT
- Basic Syntax
- Variable Selection Using Partitionning
- References
- Conclusion / Questions
Overview of Proc GLMSELECT

Performs **effect selection** in the framework of general linear models

- A variety of model selection methods are available
- Extensive capabilities for customizing the selection with a wide variety of selection and stopping criteria
- The procedure also provides graphical summaries of the selection search
- It produces output data sets and supports the SCORE statement
Model Specification

- supports different parameterizations for classification effects
- supports any degree of interaction (crossed effects) and nested effects
- supports hierarchy among effects
- supports partitioning of data into training, validation, and testing roles
- supports constructed effects including spline and multimember effects
Selection Control

- provides multiple effect selection methods
- enables selection from a very large number of effects (tens of thousands)
- 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
- supports data resampling and model averaging
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
Agenda

- Overview of GLMSELECT
- Basic Syntax
- Variable Selection Using Partitioning
- References
- Conclusion / Questions
Basic Syntax

PROC GLMSELECT <options> ;
   BY variables ;
   CLASS variable </(v-options)> <variable </(v-options ...)> </v-options> </options> ;
   EFFECT name = effect-type ( variables </options> ) ;
   FREQ variable ;
   MODEL variable = </effects> </options> ;
   MODEL AVERAGE <options> ;
   OUTPUT </OUT=SAS-data-set> <keyword =name> </keyword =name> ;
   PARTITION <options> ;
   PERFORMANCE <options> ;
   STORE </OUT=item-store-name </LABEL=label> ;
   WEIGHT variable ;
   
   SELECTION= Specifies the model selection method
Model Selection Methods

SELECTION=\texttt{method} <(\texttt{method options})>

- None
- Forward
- Backward
- Stepwise
- LAR – Least Angle Regression
- LASSO – Least Absolute Shrinkage and Selection Operator
SELECTION=method <(method options)>

<table>
<thead>
<tr>
<th>Option</th>
<th>FORWARD</th>
<th>BACKWARD</th>
<th>STEPWISE</th>
<th>LAR</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOP =</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>CHOOSE =</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>STEPS =</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>MAXSTEPS =</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>SELECT =</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCLUDE =</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SENTRY =</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLSTAY =</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DROP =</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADAPTIVE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LS COEFFS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Criteria**

- **ADJRSQ**: Adjusted R-square statistic
- **AIC**: Akaike’s information criterion
- **AICC**: Corrected Akaike’s information criterion
- **BIC**: Sawa Bayesian information criterion
- **CP**: Mallows C(p) statistic
- **CV**: Predicted residual sum of square with k-fold cross validation
- **PRESS**: Predicted residual sum of squares
- **SBC**: Schwarz Bayesian information criterion
- **VALIDATE**: Average square error for the validation data
Agenda

- Overview of GLMSELECT
- Basic Syntax
- Variable Selection Using Partitioning
- References
- Conclusion / Questions
Signal versus Noise

Predictive Modeling

- Target = Signal + Noise
- Signal = Systematic Variation = Predictable
- Noise = Random Variation = Unpredictable
Effect selection by partitioning the data

Very useful in the context of predictive modeling

- Split the data in two partitions: training & validation
- You fit (train) the model on the training partition
- The model is evaluated on the validation data
- The best model is the simplest model that has the best performance on the validation data
- The goal is to avoid overfitting
Agenda

- Overview of GLMSELECT
- Basic Syntax
- Variable Selection Using Partitionning
- References
- Conclusion / Questions
Reference Materials

- Proc GLMSELECT Documentation

- Introducing the GLMSELECT PROCEDURE for Model Selection
Conclusion

Proc GLMSELECT is a powerful tool for automatic variable selection for linear models

- Very flexible with many selection methods and stopping criterion to choose from:
  - LARS and LASSO
  - Resampling techniques: k-fold validation and bootstrap
  - Data partitionning
Questions?

THANK YOU!

Sylvain.Tremblay@sas.com