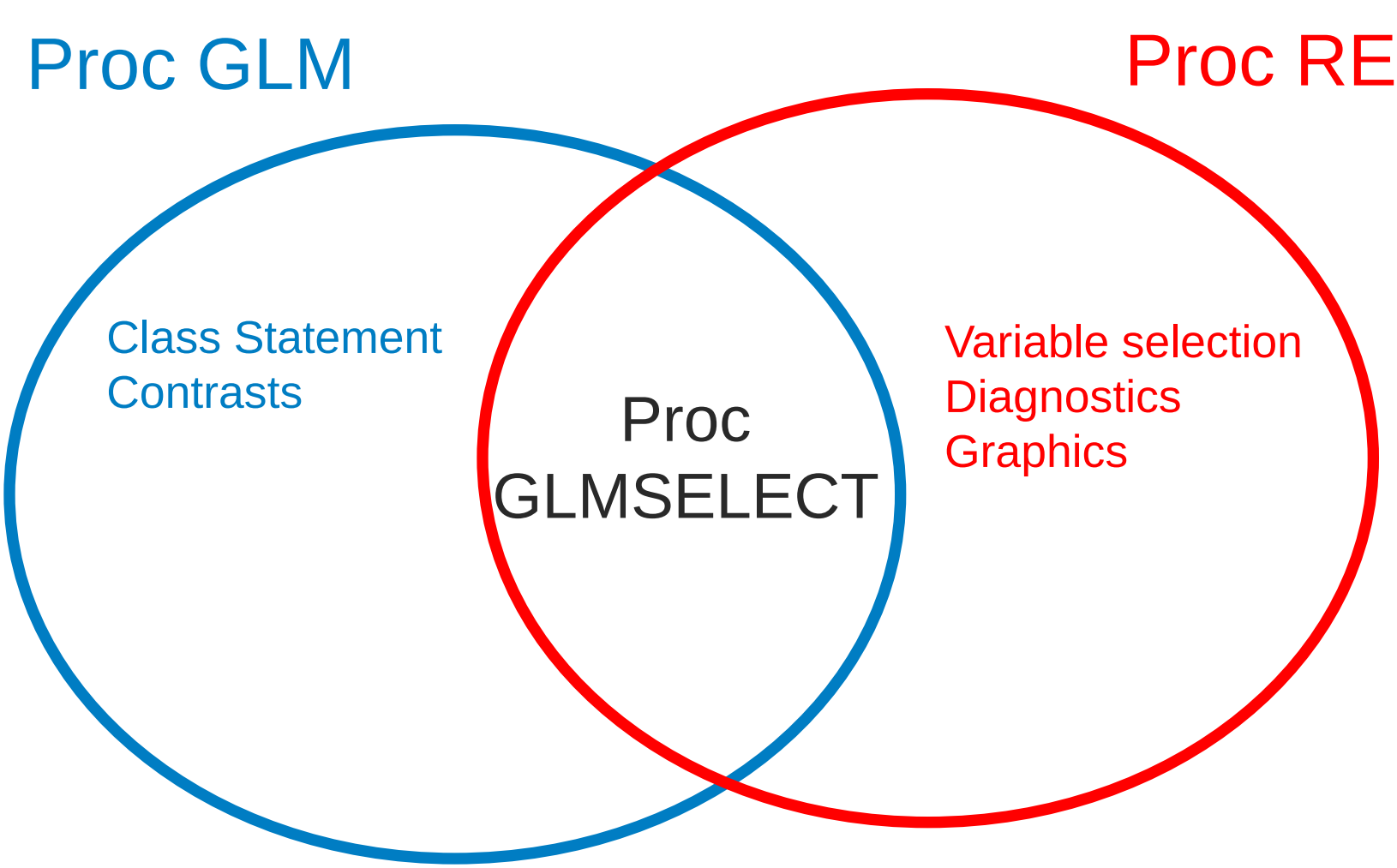
Winnipeg SAS User Group Meeting
May 11, 2012

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 Partitionning
- 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> ;
    MODELAVERAGE <options> ;
    OUTPUT <OUT=SAS-data-set> <keyword <=name> > <...keyword=name> ;
    PARTITION <options> ;
    PERFORMANCE <options> ;
    SCORE <DATA=SAS-data-set> <OUT=SAS-data-set> ;
    STORE <OUT=>item-store-name </ LABEL='label'> ;
    WEIGHT variable ;
```

SELECTION= Specifies the model selection method

Model Selection Methods

SELECTION=method <(method options)>

- None
- Forward
- Backward
- Stepwise
- LAR – Least Angle Regression
- LASSO – Least Absolute Shrinkage and Selection Operator

Model Selection Methods – Method Options

SELECTION=method <(method options)>

Option	FORWARD	BACKWARD	STEPWISE	LAR	LASSO
STOP =	x	x	x	x	
CHOOSE =	x	x	x	x	
STEPS =	x	x	x	x	
MAXSTEPS =	x	x	x	x	
SELECT =	x	x	x		
INCLUDE =	x	x			
SLENTRY =	x				
SLSTAY =		x			
DROP =					
ADAPTIVE					
LSCOEFFS					

Option	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

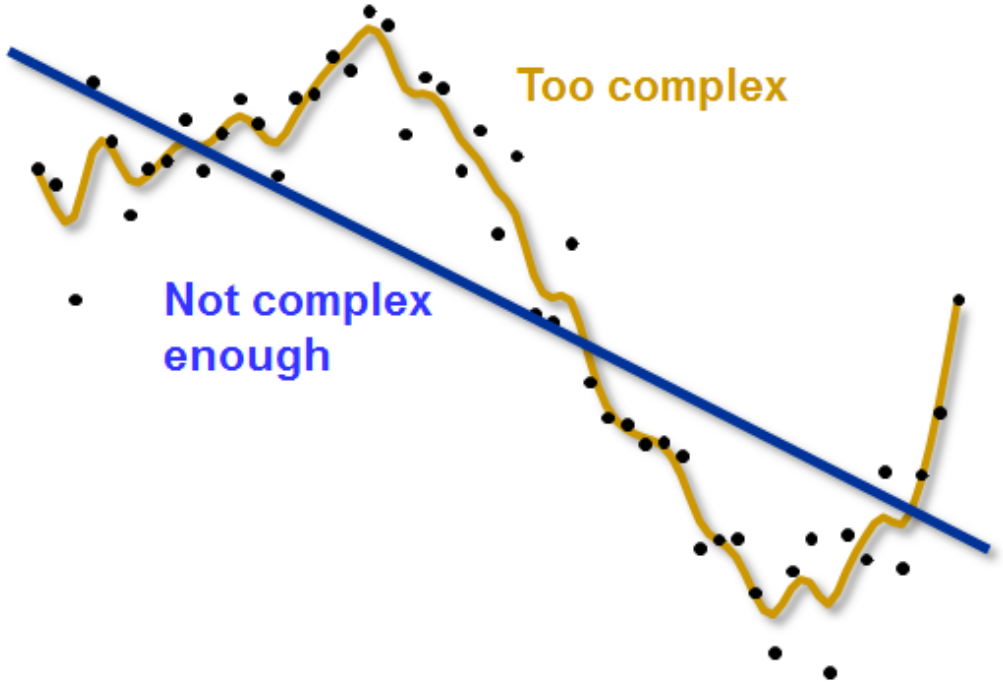
Agenda

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

Signal versus Noise

Predictive Modeling

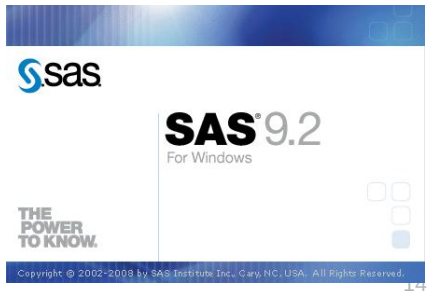
- Target = Signal + Noise
- Signal = Systematic Variation = Predictable
- Noise = Random Variation = Unpredictable



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

http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#glmselect_toc.htm

- Introducing the GLMSELECT PROCEDURE for Model Selection

http://www2.sas.com/proceedings/sugi31/207-31.pdf

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 partitionnning

Questions?

THANK YOU!

Sylvain.Tremblay@sas.com