Tips for Preparing Data for Regression Analyses

Michael Lerner
Education Statistics & Analysis Branch
Ontario Ministry of Education
March 4, 2016

Disclaimer

- The suggestions presented here are the responsibility of the author; the Ontario Ministry of Education has no responsibility and does not endorse the suggestions presented here.

Purpose

- To acquaint you with the capacities of the GLMMOD procedure and why it may be preferred in certain situations:
  - Where screening of statistical effects at a given level of significance is desired
  - Where a particular regression procedure does not support a class statement or direct modelling of interactions
  - As a 1st step in preparation of data for regression analysis

About GLMMOD

- GLMMOD creates the design matrix used in the GLM procedure that can then be used in other SAS procedures. It specifies statistical effects in same way as GLM but does not perform statistical analysis.

Example using SASHELP. CARS data set:

```
PROC GLMMOD DATA=SASHELP.CARS OUTDESIGN=X;
CLASS DRIVETRAIN TYPE ORIGIN CYLINDERS;
MODEL MSRP=TYPE DRIVETRAIN CYLINDERS HORSEPOWER CYLINDERS*HORSEPOWER;
WHERE CYLINDERS>3;
RUN;
```

- Code requests main effects for classification variables ‘type’, ‘drivetrain’ & ‘cylinders’ and for the interval variable ‘horsepower’, and then asks for the interaction effects for ‘horsepower’ by ‘cylinders’.
- The only difference is in the ‘PROC GLMMOD’ statement where the design matrix, ‘x’ is requested via ‘OUTDESIGN=X’.

What does the Design Matrix Have?

- The design matrix has the data recoded to correspond to what would be used in the corresponding GLM model.
- In our example:
  - 1st column, intercept
  - columns 2 – 7, vehicle type
  - columns 8 – 10, drive train
  - columns 11 – 16, cylinders
  - column 17, horsepower
  - columns 18 – 23, interactions of ‘horsepower’ with ‘cylinders’.
    - See table on next page (part of default output)

What does the Design Matrix Have? – Contd.

Parameter Definitions				
Column Number	Name of Associated Effect	CLASS Variable Values		
		DriveTrain	Type	Cylinders
1	Intercept			
2	Type		Hybrid	
3	Type		SUV	
4	Type		Sedan	
5	Type		Sports	
6	Type		Truck	
7	Type		Wagon	
8	DriveTrain	All		
9	DriveTrain	Front		
10	DriveTrain	Rear		
11	Cylinders			4
12	Cylinders			5
13	Cylinders			6
14	Cylinders			8
15	Cylinders			10
16	Cylinders			12
17	Horsepower			
18	Horsepower*Cylinders			4
19	Horsepower*Cylinders			5
20	Horsepower*Cylinders			6
21	Horsepower*Cylinders			8
22	Horsepower*Cylinders			10
23	Horsepower*Cylinders			12

Using the Results of GLMMOD

- A linear model using the design matrix can be estimated using for instance the REG procedure. In our example:

```
PROC REG DATA=X;
MODEL MSRP=COL2-COL6 COL8-COL9 COL11-COL15 COL17 COL19-COL22;
RUN;
```

- The columns corresponding to the last levels of a classification variable and of the interactions of ‘horsepower’ with ‘cylinders’ have been omitted. Had no intercept been specified in GLMMOD, then all effects could be entered.

Conclusion – So why use GLMMOD?

- The example that has been shown is trivial – it could be arrived at by using GLM.
- You could also use the GLMSELECT procedure with the model statement option of 'HIERARCHY=NONE' to screen the effect of predictors using a wide array of criteria.
- But if GLMSELECT does not meet your needs for screening effects due to the nature of the outcome variable, then GLMMOD is a necessary 1st step.
  - Example: specify a GLMMOD design matrix for the cars data set where type of car is sedan or not with 'drivetrain', 'cylinders' & 'horsepower' as predictors, and then use logistic regression to predict the probability of a sedan being observed.
    - Classification level predictors such as 'drivetrain' do not go into the 'Class' statement as they are already 'dummied'.
- The procedure is also useful:
  - when you are using a regression procedure that does not have a class statement (e.g., the NLIN procedure that estimates non-linear models or the SYSLIN procedure for systems of equations)
  - as a 1st step for further preparation of data for regression. If desirable, 'horsepower' and each interaction of 'horsepower' with 'number of cylinders' could have been centered around their respective means using the design matrix data in the STDIZE procedure
- And this can be done without the added effort of a data step and the enhanced possibility of errors.

Additional Resources

- Rick Wicklin's blog 'The DO Loop'
- February 22nd post: 'Create Dummy Variables in SAS'
- February 24th post: 'Four Ways to Create a Design Matrix in SAS'