How can the PROC GLMSELECT statement be used in SAS?

Authored by

stats writer

June 23, 2024

The PROC GLMSELECT statement is a powerful tool in SAS that allows users to perform regression analysis and select the optimal set of variables for a given model. This statement helps to automate the variable selection process and provides various options for model building and diagnostics. It can be used to fit linear, logistic, and multinomial regression models and allows for the inclusion of categorical and continuous variables. Additionally, the PROC GLMSELECT statement offers advanced features such as stepwise selection, forward and backward selection, and model comparison using various criteria. Overall, this statement is essential for efficient and accurate model building and selection in SAS.

Use the PROC GLMSELECT Statement in SAS

You can use the PROC GLMSELECT statement in SAS to select the best regression model based on a list of potential predictor variables.

The following example shows how to use this statement in practice.

Example: How to Use PROC GLMSELECT in SAS for Model Selection

Suppose we want to fit a multiple linear regression model that uses (1) number of hours spent studying, (2) number of prep exams taken and (3) gender to predict the final exam score of students.

First, we'll use the following code to create a dataset that contains this information for 20 students:

```
/*create dataset*/
data exam_data;
input hours prep_exams gender $ score;
datalines;
1 1 0 76
2 3 1 78
2 3 0 85
4 5 0 88
2 2 0 72
1 2 1 69
5 1 1 94
4 1 0 94
2 0 1 88
4 3 0 92
4 4 1 90
3 3 1 75
6 2 1 96
5 4 0 90
3 4 0 82
4 4 1 85
6 5 1 99
2 1 0 83
1 0 1 62
2 1 0 76
```

```
;
run;

/*view dataset*/proc printdata=exam_data;
```

Obs	hours	prep_exams	gender	score
1	1	1	0	76
2	2	3	1	78
3	2	3	0	85
4	4	5	0	88
5	2	2	0	72
6	1	2	1	69
7	5	1	1	94
8	4	1	0	94
9	2	0	1	88
10	4	3	0	92
11	4	4	1	90
12	3	3	1	75
13	6	2	1	96
14	5	4	0	90
15	3	4	0	82
16	4	4	1	85
17	6	5	1	99
18	2	1	0	83
19	1	0	1	62
20	2	1	0	76

Next, we'll use the PROC GLMSELECT statement to identify the subset of predictor variables that produces the best regression model:

```
/*perform model selection*/
proc glmselectdata=exam_data;
class gender;
model score = hours prep_exams gender;
run;
```

Note: We included gender in the class statement because it is a categorical variable.

The first group of tables in the output shows an overview of the GLMSELECT procedure:

The GLMSELECT Procedure

Data Set	WORK.EXAM_DATA
Dependent Variable	score
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None

Number of Observations Read	20
Number of Observations Used	20

Class Level Information		
Class	**Levels**	**Values**
gender	2	0 1

Dimensions	
Number of Effects	4
Number of Parameters	5

We can see that the criterion used to stop adding or removing variables from the model was SBC, which is *Schwarz Information Criterion*, sometimes called the *Bayesian Information Criterion*.

Essentially the PROC GLMSELECT statement keeps adding or removing variables from the model until it finds the model with the lowest SBC value, which is considered the "best" model.

The next group of tables shows how the stepwise

selection stopped:

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	SBC
0	Intercept		1	1	93.4337
1	hours		2	2	70.4452*

* Optimal Value of Criterion

Selection stopped at a local minimum of the SBC criterion.

Stop Details

Candidate For	Effect	Candidate SBC		Compare SBC
Entry	gender	71.7383	>	70.4452
Removal	hours	93.4337	>	70.4452

We can see that a model with only the intercept term had a SBC value of 93.4337.

The next best possible way to improve the model was to add gender as a predictor variable, but this actually increased the SBC value to 71.7383.

Thus, the final model only includes the intercept term and hours studied.

The last portion of the output shows the summary of this fitted regression model:

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept hours

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	1338.29063	1338.29063	48.00
Error	18	501.90937	27.88385	
Corrected Total	19	1840.20000		

Root MSE	5.28052
Dependent Mean	83.70000
R-Square	0.7273
Adj R-Sq	0.7121
AIC	90.45375
AICC	91.95375
SBC	70.44521

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	67.161689	2.663270	25.22
hours	1	5.250257	0.757847	6.93

We can use the values from the Parameter Estimates table to write the fitted regression model:

Exam Score = 67.161689 + 5.250257(hours studied)

We can also see various metrics that tell us how well this model fits the data:

The R-Square value tells us the percentage of variation in the exam scores that can be explained by the number of hours studied and the number of prep exams taken.

In this case, 72.73% of the variation in exam scores can be explained by the number of hours studied and number of prep exams taken.

The Root MSE value is also useful to know. This represents the average distance that the observed values fall from the regression line.

In this regression model, the observed values fall an average of 5.28052 units from the regression line.

Note: Refer to the for a complete list of potential arguments you can use with PROC GLMSELECT.

The following tutorials explain how to perform other common tasks in SAS: