SESUG Paper 40 - 2025

Modernizing Predictive Modeling in SAS®: When to Move Beyond PROC GLM

Duha Khan, Dr. Jonathan Duggins, North Carolina State University

ABSTRACT

SAS® offers many procedures that can be used to conduct predictive analysis. PROC GLM is one of the older tools used to fit general linear models to datasets. While it includes many options that allow for the implementation of classes, subgroups, weighting, and more, there are times when another, newer procedure may be preferred. For example, since PROC GLM is not a high-performance procedure, it is sometimes better to use a high-performance alternative like PROC HPREG, which is better suited for larger datasets. Additionally, PROC GLM requires manual variable selection from the programmer. While this can be useful at times, there are other cases where the automated selection with PROC GLMSELECT can prove to be far more efficient. This paper explores the pros and cons of these procedures with examples to help users choose the most appropriate predictive modeling tool for real-world applications.

INTRODUCTION

In the field of statistical analysis and predictive modeling, building effective linear models is a foundational skill for analysts across many industries. This paper explores how PROC GLM, PROC GLMSELECT, PROC REG, and PROC HPREG can be used to build linear models for predicting a quantitative outcome. While each of these procedures shares the same fundamental goal, they offer different options that make them better suited for specific modeling tasks or data environments.

The primary goal of this paper is to compare these procedures with a focus on variable selection and computational efficiency. This will provide insight on how to choose the appropriate procedure for a SAS user's data analysis needs, especially when working with large datasets, correlated predictors, or when prediction accuracy is critical.

The example code provided in the following sections uses a dataset containing twenty-one potential predictors for human life expectancy, as well as a variable representing life expectancy as reported by the World Health Organization (Kumar, JARSHI. "Life Expectancy (WHO)").

FITTING A MODEL VIA PROC GLM

The Generalized Linear Model (GLM) procedure uses the least squares method to fit general linear models to data. This allows one to examine the relationships between variables and produce predictions for desired variables based on chosen predictors.

To demonstrate the basic functionality of PROC GLM, the following example produces a model of life expectancy as a function of five health and demographic variables manually selected by the programmer:

Example 1

```
PROC GLM DATA=MyData.life;
    ① MODEL life_expectancy = bmi hiv_aids alcohol polio population
            / ②ss3;
    ③ OUTPUT OUT=SESUG.predictions1 ④ PREDICTED=life_prediction;
RUN;
QUIT;
```

① The MODEL statement contains the response variable(s) to the left of the equal sign with the predictor variable(s) to the right. In this case, the goal is to examine life expectancy in terms of BMI, HIV/AIDS, alcohol, polio, and population.

② The SS3 option tells SAS to display only the type III sums of squares for each parameter. The default is to display type I and type III sums of squares.

③ The OUTPUT statement allows the user to create a new dataset that includes values that are calculated after fitting the model.

④ The PREDICTED option was used to create a new variable called life_prediction. This variable contains the predicted life expectancy for a person based on the predictors included in this model.

- This dataset already contains the life_expectancy variable with the observed values of life expectancy. The observed values and the predicted values can be used to calculate residuals and get a better idea of the magnitude of the prediction error in the linear model.

However, PROC GLM depends on manual variable selection. This can be a useful feature when building upon previous findings in a research application, when there is prior knowledge about which variables are particularly of interest as predictors. Unfortunately, in some cases, manual variable selection can introduce researcher bias. When beginning analysis to find relationships in data that there is no prior knowledge on, it may be tempting to list all the available variables in the model statement and work from there. While this may work if there are only a few predictors, too many predictor variables in a linear model lead to the risk of overfitting the data. Overfitting occurs when the model you produce is too specific to the data you derived it from. Such a model might seemingly perform well on that specific dataset, but it doesn't accurately "generalize" as desired. This can impact the output associated with each of the listed predictors in the MODEL statement and possibly decrease the overall predictive accuracy of the model (Chen, Yun-chun, et al. 2013).

Another issue that programmers may have to deal with is shared variance. This is a phenomenon that occurs when the selected predictors are not independent. When certain predictors are correlated with each other, it becomes difficult to isolate which is the strongest predictor out of them to include in the model. This can lead to inaccurate p-values and an overall poor outcome.

Example 2

```
PROC GLM DATA=MyData.life;
    MODEL life_expectancy = bmi hiv_aids thinness_1_19_years
          thinness_5_9_years adult_mortality alcohol diphtheria gdp
          hepatitis_b income_composition_of_resources measles polio
          population schooling total_expenditure year infant_deaths
          percentage_expenditure under_five_deaths / ss3;
RUN;
QUIT;
```

Such a case can be seen in Example 2 with the analysis of the life expectancy dataset. When PROC GLM is run with just five variables in the MODEL statement (Example 1, .

Table 1), the predictor polio is found to be statistically significant. When GLM is run again with eighteen other predictors (Example2,

Table 2), the p-values associated with polio are found to be much higher– and insignificant. Seeing that polio gets "drowned out" in the larger model suggests that it shares variance with some of the other listed variables. It is possible that polio is an important predictor being overshadowed by others. Another possibility is that polio is in fact not significant but only appeared so in the simpler model due to chance. Either way, when a situation like this arises, steps should be taken to clarify such overlap in variables to produce the most optimal predictive model possible. Since this analysis is exploratory in nature, it may be preferable to utilize a procedure that facilitates an automatic variable selection method, such as PROC GLMSELECT.

Table 1: Output From Example 1

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bmi	1	17945.60592	17945.60592	514.58	<.0001
hiv_aids	1	37946.21407	37946.21407	1088.08	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Alcohol	1	7584.46442	7584.46442	217.48	<.0001
Polio	1	8904.20174	8904.20174	255.32	<.0001
Population	1	6.34319	6.34319	0.18	0.6698

Table 2: Output From Example 2

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bmi	1	355.109104	355.109104	27.98	<.0001
hiv_aids	1	8073.582446	8073.582446	636.13	<.0001
thinness_1_19_years	1	0.023854	0.023854	0.00	0.9654
thinness_5_9_years	1	13.270305	13.270305	1.05	0.3067
adult_mortality	1	3864.252464	3864.252464	304.47	<.0001
Alcohol	1	125.091202	125.091202	9.86	0.0017
Diphtheria	1	67.200020	67.200020	5.29	0.0215
GDP	1	13.838764	13.838764	1.09	0.2965
hepatitis_b	1	3.482249	3.482249	0.27	0.6005
income_composition_o	1	1999.342494	1999.342494	157.53	<.0001
Measles	1	13.547246	13.547246	1.07	0.3017
Polio	1	15.457918	15.457918	1.22	0.2699
Population	1	1.793460	1.793460	0.14	0.7070
Schooling	1	2989.597516	2989.597516	235.56	<.0001
total_expenditure	1	71.574456	71.574456	5.64	0.0177
Year	1	401.133063	401.133063	31.61	<.0001
infant_deaths	1	888.629000	888.629000	70.02	<.0001
percentage_expenditu	1	38.142449	38.142449	3.01	0.0832
under_five_deaths	1	954.155447	954.155447	75.18	<.0001

This output is a great example of showing multicollinearity, which occurs when predictors in a model are highly correlated with each other. As a result, it becomes difficult to isolate each variable's individual effect on the response and makes interpretation difficult. However, this issue is more detrimental to statistical inference than the prediction values. Still, when unsure, it is a good idea to use the CORR procedure to quickly check for correlation between variables.

FITTING A MODEL VIA PROC GLMSELECT

PROC GLMSELECT is a specialized version of PROC GLM that is designed for automatic model building and variable selection.

The following code shows a similar analysis as done above, but this time with GLMSELECT:

Example 3

```
PROC GLMSELECT DATA=MyData.life;
    ① MODEL life_expectancy = bmi hiv_aids thinness_1_19_years
             thinness_5_9_years adult_mortality alcohol diphtheria gdp
             hepatitis_b income_composition_of_resources measles polio
             population schooling total_expenditure year infant_deaths
             percentage_expenditure under_five_deaths / ② SHOWPVALUES
                                                        ③ SELECTION=lasso;
    ④ OUTPUT OUT=SESUG.predictions2 PREDICTED=life_prediction;
RUN;
QUIT;
```

① This time, all the variables from the dataset were included in the MODEL statement

② The SHOWPVALUES option is included to display p-values in the "ANOVA" and "Parameter Estimates" tables.

③ The SELECTION method chosen is lasso (least absolute shrinkage and selection operator). This method is preferable when predictors in a model have a shared variance (Tibshirani, Robert. 1996)

④ The life expectancy predictions from this model were OUTPUT to a new file called *SESUG.predictions2*.

As seen in Table 3, after the automatic selection process done by SAS, polio was not included in the model as a predictor for life expectancy. GLMSELECT runs by checking the fit of many models based on the selection method (in this case, LASSO) and outputs the model that has the best performance based on a certain criterion (which can either be specified, or has defaults based on the selection method). In this example, polio was not found to have a significant unique contribution to predicting life expectancy. This case is a great example of how PROC GLMSELECT can drop the redundant or less useful variables from a model.

Table 3: Output From Example 3

Parameter Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	54.067967
bmi	1	0.027993
hiv_aids	1	-0.395107
thinness_5_9_years	1	-0.020111
adult_mortality	1	-0.017752
Diphtheria	1	0.008960
GDP	1	0.000014409
income_composition_o	1	9.786309
Polio	1	0.000636
Schooling	1	0.894910
percentage_expenditu	1	0.000174

FITTING A MODEL VIA PROC REG

The REG procedure allows a programmer to fit standard linear regression models with manually chosen numeric predictors. The output includes parameter estimates and statistics about how the model fits. The following code is a very basic example utilizing this procedure:

Example 4

```
PROC REG DATA=MyData.life;
    ① MODEL life_expectancy = bmi hiv_aids alcohol polio population;
    ① OUTPUT OUT=SESUG.predictions3 PREDICTED=life_prediction;
RUN;
```

① The MODEL and OUTPUT statements function just how they do in PROC GLM.

There is some overlap between PROC REG and PROC GLM. The file *SESUG.predictions3* is found to be the exact same as *SESUG.predictions1*. Although the layout differs slightly, both GLM (Example 1) and REG (Example 4) produce identical ANOVA tables. Further details on the differences between PROC GLM and PROC REG are provided in the 'Comparisons' section of this paper.

FITTING A MODEL VIA PROC HPREG

Now, in addition to PROC REG, SAS also offers users a high-performance version of the procedure, appropriately named PROC HPREG. Similar to PROC GLMSELECT, the HPREG procedure conducts automatic variable selection:

Example 5

```
PROC HPREG DATA=MyData.life;
    MODEL life_expectancy = bmi hiv_aids thinness_1_19_years
          thinness_5_9_years adult_mortality alcohol diphtheria gdp
          hepatitis_b income_composition_of_resources measles polio
          population schooling total_expenditure year infant_deaths
          percentage_expenditure under_five_deaths;
    ① SELECTION METHOD = LASSO;
    OUTPUT OUT=SESUG.predictions4 PREDICTED=life_prediction;
RUN;
```

① PROC HPREG allows the option to specify the SELECTION METHOD. For consistency, LASSO was used again with the same predictors as specified in the GLMSELECT procedure (Example 3).

The PARTITION option can also be used to split the dataset into three disjoint sets. This way, the model can be trained, validated, and tested simultaneously. The training set includes the data that the model will "learn" from. The validation set is a different portion of the data that checks to see how well the model performs with data it hasn't seen before. Then, the testing set does a "final test" of the model with another set of unseen data to do a final evaluation of how the model will perform on new data. This is a feature that is incredibly useful during the process of predictive modeling.

The most notable fact about the way the HPREG procedure runs is the fact that it carries out multithreaded processing. This means that the procedure is able to split the work it needs to do across multiple CPU cores so that the data can be processed far more efficiently.

COMPARISONS

A COMPARISON OF PROC GLM AND PROC GLMSELECT

It may be beneficial to examine the difference in the prediction accuracy of the different linear models created in the previous sections. Using the output *SESUG.predictions1* from Example 1 with 5 manually selected variables, a DATA step is used followed by PROC MEANS to calculate the average residual in the life expectancy value in the data versus the life expectancy value predicted from the generated model. The same process is repeated with the *SESUG.predictions2* output produced from the GLMSELECT

procedure in Example 3. It was found that the average residual life expectancy from the GLM model was 4.52 years, and the average residual from the GLMSELECT model was 3.01 years.

Choosing between GLM and GLMSELECT depends on what the programmer already knows going into the analysis. When there is prior knowledge that can aid the manual variable selection method, PROC GLM would be sufficient. However in situations where little is known beforehand and variable selection is necessary, GLMSELECT may be the wiser choice.

A COMPARISON OF PROC GLM AND PROC REG

There are some key differences between REG and GLM. Most notably, PROC GLM includes the CLASS option, which allows for the handling of categorical variables. The REG procedure doesn't have a similar option in place, thus making the inclusion of categorical variables much more difficult. If one wants to include categorical variables as predictors, they may want to turn towards PROC GLM or PROC HPREG instead.

Another difference is that PROC REG can handle multiple MODEL statements within one step and will produce the models corresponding to all the statements in the output. GLM, on the other hand, does not have this capability (Koontz, Stephen R. 1999.). So, if more than one model needs to be produced with all continuous variables, the REG procedure may be preferred.

A COMPARISON OF PROC REG AND PROC HPREG

Unlike REG, HPREG offers the CLASS option to handle categorical variables as predictors. Thus, in a case that one would prefer to use REG, but also need this option, the HPREG procedure is available. Additionally, the PARTITION option mentioned earlier is another high-performance perk of PROC HPREG.

Additionally, while PROC REG does support the use of some selection methods (such as FORWARD, BACKWARD, and STEPWISE), it is rather limited in comparison to the newer procedures. PROC HPREG supports more of the modern methods, such as LASSO.

Besides differences in options, PROC HPREG also differs computationally. While PROC REG runs in a single-threaded environment, PROC HPREG uses multi-threaded processing, which makes it optimal for larger datasets. Of course, this is only possible if the machine has multiple CPU cores available. While this is rarely an issue with modern computers, it's worth noting that without multiple cores, PROC HPREG will offer little to no efficiency advantage over PROC REG.

```
NOTE: PROCEDURE REG used (Total process time):
      real time           0.81 seconds
      user cpu time       0.33 seconds
```

Figure 1

```
NOTE: The HPREG procedure is executing in single-machine mode.
NOTE: There were 2938 observations read from the data set MYDATA.LIFE.
NOTE: The data set SESUG.PREDICTIONS4 has 2938 observations and 1 variables.
NOTE: PROCEDURE HPREG used (Total process time):
      real time           0.08 seconds
      user cpu time       0.06 seconds
```

Figure 2

The log points out that the procedure is executing single machine mode. That's because the number of processors weren't specified in the PERFORMANCE statement with the THREADS= option in the code. So, although it looks like PROC HPREG ran a bit faster here, it's not because the data was split up across multiple CPU cores. Other times, SAS detects that single-machine mode is sufficient to carry out the desired analysis. The main takeaway here is to ensure that the correct syntax is being used in order

to use the full capabilities of this high-performance procedure.

CONCLUSION

This paper introduced multiple statistical modeling procedures in SAS, clarifying how PROCs often used for predictive modeling align and differ in the context of general linear models and predictive analysis.

One of the key takeaways is about how the PROC choice depends on both the goals of the analysis as well as the available knowledge. For example, manual variable selection in PROC REG or GLM can be ideal when the model is already well-defined, while automated selection in PROC GLMSELECT or HPREG offers efficiency in exploratory settings. The high-performance capabilities of PROC HPREG allow multithreading for large datasets, although the advantages diminish on single-core machines. Transitioning between PROCs is easier when their syntax and capabilities are viewed side-by-side, thereby reducing the learning curve for new methods.

This information aids new users in choosing the most appropriate procedure and helps experienced users transition smoothly between them.

REFERENCES

Chen, Yun-chun, et al. 2013. “The Predictive Value of Body Mass Index on the Survival of Patients with Stage IV Non-small Cell Lung Cancer: A Retrospective Analysis of 1,494 Cases.” Journal of Thoracic Disease, 5(6): 841–848. https://jtd.amegroups.org/article/view/3081/3566.

DeVille, Tom. 2017. “An Introduction to Generalized Linear Models with Examples in SAS®.” Proceedings of the SAS Global Forum 2017 Conference, Orlando, FL: SAS Institute Inc. Available at https://support.sas.com/resources/papers/proceedings17/1404-2017.pdf.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. “The Elements of Statistical Learning.” Stanford University. Accessed August 3, 2025. https://web.stanford.edu/~hastie/ElemStatLearn/.

Koontz, Stephen R. 1999. “Econometric Models for Count Data.” Proceedings of the MidWest SAS Users Group Conference, Kansas City, MO: MidWest SAS Users Group. Available at https://www.lexjansen.com/mwsug/1999/paper15.pdf.

Kumar, JARSHI. “Life Expectancy (WHO).” Kaggle. Accessed August 3, 2025. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who.

SAS Institute Inc. “GLM Procedure Overview.” SAS Documentation. Accessed August 3, 2025. https://documentation.sas.com/doc/en/statug/15.2/statug_glm_overview.htm.

SAS Institute Inc. “GLM Procedure: Syntax.” SAS Documentation. Accessed August 3, 2025. https://documentation.sas.com/doc/en/statug/15.2/statug_glm_syntax17.htm.

SAS Institute Inc. “GLMSELECT Procedure: Selection Details.” SAS Documentation. Accessed August 3, 2025. https://documentation.sas.com/doc/en/statug/15.2/statug_glmselect_details10.htm.

SAS Institute Inc. “HPREG procedure: Syntax.” SAS Documentation. Accessed August 3, 2025. https://documentation.sas.com/doc/en/statug/15.3/statug_hpreg_syntax09.htm.

SAS Institute Inc. “System Options: LESYSOPT Reference.” SAS Documentation. Accessed August 3, 2025. https://documentation.sas.com/doc/en/vdmmlcdc/8.1/lesysoptsref/n149q4yzzknwqbn1obgqbfrc1x0n.htm.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” Journal of the Royal Statistical Society: Series B (Methodological), 58(1): 267–288. Available at https://webdoc.agsci.colostate.edu/koontz/arec-econ535/papers/Tibshirani%20(JRSS-B%201996).pdf.

UCLA Institute for Digital Research and Education. “GLM Output in SAS.” UCLA Statistical Consulting. Accessed August 3, 2025. https://stats.oarc.ucla.edu/sas/output/glm/.

UCLA Institute for Digital Research and Education. “Regression Analysis Output in SAS.” UCLA Statistical Consulting. Accessed August 3, 2025. https://stats.oarc.ucla.edu/sas/output/regression-analysis/.

Worcester Polytechnic Institute. “Chapter 30: The GLM Procedure.” SAS/STAT User’s Guide. Accessed August 3, 2025. http://www.math.wpi.edu/saspdf/stat/chap30.pdf.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Dr. Jonathan Duggins for his guidance and mentorship throughout the development of this paper. His feedback and encouragement were invaluable in helping me shape and refine my work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Duha Khan
duhathekhan@gmail.com