REG, GLM, ANOVA: Which one? Why? How?

Linda M. Quinn
QED Industries and Cleveland State University

ABSTRACT

Modeling the relationship between a response variable and one or more independent variables can be done within the SAS® System using many different PROCs including REG, ANOVA, and GLM. This beginning level tutorial will show which procedure is the best choice under a variety of different conditions, why one might be a better choice than another, and the difference in output. Examples and coding will be provided.

KEYWORDS

REG, ANOVA, GLM, analysis of variance, regression

INTRODUCTION

The three procedures, REG, ANOVA, and GLM, are sometimes used interchangeably. The purpose of this presentation is to discuss the relative advantages of one procedure over the other. Differences exist in the computing efficiencies, diagnostic options, and specification of interaction terms.

EXAMPLE

Here is an example dataset. It will be used to show the different results produced results running the different procedures. The dataset consists of a continuous outcome variable, value of a Honda Civic, and several independent variables: age of car (continuos in years), transmission type (dichotomous - automatic or standard), number of doors (dichotomous - 2 or 4), and mileage (continuous in miles).

There are several questions that can be asked. Do any of the independent variables help explain car value? Can any of the independent variables predict the car's value? Can a model be developed to explain or predict a car's value from the data collected?

So the dependent variable is VALUE. The independent variables are AGE, TRANSM, DOORS, MILES.

Some characteristics to remember are: TRANSM is the only character variable; and both TRANSM and DOORS are dichotomous.

Let's start by looking at each of the three procedures in their most basic form and what they can do for this example.

PROC REG

The REG procedure is used to fit ordinary least squares (OLS) regression models. REG is a general purpose regression procedure.

REG will not accept a classification variable. If we want to model VALUE using TRANSM, we need to create an indicator variable: AUTO equals 1 if automatic and 0 if standard transmission. If TRANSM would have more than two levels, a series of indicator variables would need to be created.

We can now run each independent variable in a separate model or run as a multivariable model. In REG, multiple model statements can be used, each with a label if desired.

```
proc reg;
  model value = miles;
  model value = doors;
  model value = auto;
  model value = age;
  MULTI: model value = miles doors auto age;
run;
```

Each model will produce a similar type output (see Appendix A). This will include a standard ANOVA table, a section with each parameter estimate, and its significance.

PROC ANOVA

The ANOVA procedure is used to fit analysis of variance, multivariate analysis of variance, and repeated measures analysis of variance. ANOVA is best used when the design is well balanced.

ANOVA uses a CLASS statement to specify classification variables such as DOORS and TRANSM. However, it cannot handle continuous variables such as age and mileage, even if they are covariates in an analysis of covariance. So the only predictors we could look at are DOORS and TRANSM (or AUTO will work as well). Only one model can be specified in each proc step.

```
proc anova;
  class auto;
  model value = auto;
run;
```

Looking at the model for TRANSM, the general ANOVA table (see Appendix A) has identical results even with the same labeling. The layout is slightly different. There is no parameter section but instead there is a Sums of Squares section. The p-value is identical to that from REG, despite that it is from an F-test and REG is from a t-test. (Note: this is an expected result since there is a mathematical relationship between the t and F tests.)

PROC GLM

The GLM procedure is used to fit general linear models.

GLM has a CLASS statement to specify categorical predictors. However, you can only specify one model statement per procedure.

```
proc glm;
  class auto;
  model value = auto;
run;
```

The output is very similar to that produced by ANOVA (see Appendix A). The Sums of Square section now has more options though (Types I and III).

FIRST GLANCE COMPARISONS

First, let's examine what the three procedures have it common.

- While not seen with just printed output, all three procedures are interactive. After a model statement and run statement are issued, each procedure has a variety of statements that can be executed without reinvoking the procedure. Each will produce output and still give the message "Proc REG running" until another proc or datastep in encountered, or a QUIT statement is executed.
- All three procedures produce identical analysis of variance tables for the general model, including the root mean square error, coefficient of

variation, and R-square. Each has only a slightly different layout.

- All three procedures use a similar structure of the MODEL statement. *model dependent = independents;*
- For univariate models, all three procedures give the same significance value for the independent variable.
- While not seen in this basic example, all three procedures can do multivariable modeling.

Now let's look at the differences.

- Multiple models can be run within one REG procedure and cannot with ANOVA and GLM.
- ANOVA only handles categorical independent variables; REG needs indicator variables created for categorical data. GLM can handle both types of variables. ANOVA and GLM have CLASS statements available to specify categorical variables.
- In this basic example, ANOVA and GLM produce Sums of Squares for each predictor variable and REG produces parameter estimates.

There are two main types of analyses going on in these procedures, regression and analysis of variance. Regression analyses produce a regression or prediction model with estimates for the parameters of the linear equation. Therefore it is important that categorical variables are treated as series' of indicator variables. Analysis of variance is comparing the amount of variance explained by different factors, typically categorical in nature.

The GLM procedure is a mixture of both regression and analysis of variance, called general linear models and is the most general of the analysis of variance procedures. GLM can be a real workhorse for analysis. GLM is a powerful procedure, and many times is a great substitute for both the REG procedure and the ANOVA procedure.

GLM ANALYSES

- Ordinary least squares regression, simple and multiple
- Analysis of variance, balanced and unbalanced designs
- Analysis of covariance (both continuous and categorical independent varaiables)
- Multiple analysis of variance (MANOVA)
- Specialized regressions such as weighted and polynomial regression
- Repeated measures analysis of variance

GLM VERSUS REG

Remember that the main difference between REG and GLM is that GLM didn't produce parameter estimates and couldn't run multiple model statements. There is nothing that can be done about the multiple models; however, GLM can produce parameter estimates. If there is no CLASS statement within the procedure, GLM is assuming that all the independent variables are continuous and that the analysis of interest is regression. In this case, GLM produces the parameter estimates. If there is a mixture of categorical and continuous variables or only class variables, using the SOLUTION option on the model statement will give the parameter solution to the normal equations.

```
proc glm;
  class auto;
  model value = auto/solution;
run;
```

The solution is to the normal equations and therefore there is a parameter estimate for each level of the categorical variable leading to what appears to be an overspecified model with biased estimates for the parameters.

So besides multiple model statements, are there advantages to using REG over GLM? Yes, the largest and most important advantage is that REG is designed with many regression diagnostics, particularly for the detection of collinearity, not available in GLM. GLM has some diagnostics, but the REG procedure is more exhaustive. REG can also:

- Produce partial regression leverage plots.
- Can use corrrelations or cross products as input as well as a SAS dataset.
- Provide nine different methods of model selection, including some iterative methods (stepwise, forward and backward elimination, etc.).
- Test linear and multivariate hypotheses.
- Perform weighted regression analyses.

GLM VERSUS ANOVA

So far, in the basic case, the two procedures look identical except for a few more lines with two different types of sums of squares for GLM. The main differences between the two procedures are not obvious.

- ANOVA should only be used when the analysis is based on a complete and balanced design, i.e., every combination of levels in independent variables has the same number of observations. Otherwise, ANOVA may not give accurate results and it is up to the user to validate the results. This is probably the most important difference and the one that causes most users to use GLM over ANOVA routinely.
- GLM, on the other hand, is computationally more intense if the balanced conditions of the ANOVA procedure are met. The calculations for analysis of variance are greatly simplified if the design is balanced. If computational efficiency is not an issue, use GLM over ANOVA in all cases.

Given their main differences, both can be used to run more complex ANOVA designs, such as repeated measures, nested and crossed designs. Both are also able to give post hoc multiple comparison tests.

OTHER DIFFERENCES

The other difference between the procedures is the specification of interaction terms in the model. Both ANOVA and GLM allow for shorthand notation to be used. REG requires the variables exist prior to being specified in the model. This means that the datasets must have all the possible interaction terms you are interested in.

The following statements show the analysis of value of a car based on miles driven, transmission type, and the interaction between the two.

In GLM and ANOVA:

```
model value = miles*auto;
```

In REG, you would have needed to create a third variable within a dataset (mileauto = miles*auto;) and then use it in the model statement:

```
model value = miles auto mileauto;
```

The * shorthand tells SAS to create the interaction between those two variables as well as the main effects for those variables. The | notation between variables (typically more than two) tells SAS to create all the main effects and all possible interaction terms. For example, the following statements all produce equivalent models. Notice this model includes a three variable interaction.

```
model value = miles|auto|doors;
model value = miles auto doors miles*auto
    miles*doors auto*doors miles*auto*doors;
```

FINAL NOTES

All the models share features developed throughout all the SAS procedures such as outputting results, creating new datasets, producing predictive values, and handling no-intercept models.

There are many other procedures, some with specialized purposes, that can be used for regression analysis and analysis of variance. Some of them are:

CALIS	Used for structural equation modeling and path analysis.
CATMOD	Fits linear models to categorical data.
LIFEREG	Used for failure time regression analysis.
LOGISTIC	Used for regression when the outcome variable is categorical.
MIXED	Used for mixed model development and analysis.
NLIN	Models nonlinear regression models.
NESTED	Models nested ANOVA designs,

Users should investigate the applicability of the MIXED procedure in their analyses. MIXED has features specific to mixed models that are more applicable than GLM. MIXED also has the additional feature of the Output Delivery System (ODS).

As a rule of thumb, I usually use REG for regression and GLM for analysis of variance or covariance. I usually forget to produce interaction terms and quadratic effects until I get the error message that REG doesn't recognize the shorthand. When running GLM, I usually forget that it cannot run multiple models in one invocation. The first error message usually sets me straight.

I rarely use ANOVA, and when I do, I usually run GLM as well. I am starting to run more MIXED analyses.

REFERENCES

SAS Institute, Inc. (1990), *SAS/STAT User's Guide, Version 6, Fourth Edition, Volumes 1 and 2*, Cary, NC: SAS Institute, Inc.

Little, Ramon C., Freund, Rudolf J., Spector, Philip C., *SAS System for Linear Models, Third Edition*, Cary, NC: SAS Institute Inc., 1991.

All applicable updates to the SAS/STAT User's Guide.

PRODUCT INFORMATION

AUTHOR CONTACT

Linda M. Quinn
QED Industries
1286 Gary Blvd
Brunswick, OH 44212-2912

Phone: 330-225-4184
Fax: 330-220-9036
e-mail: lmq2@po.cwru.edu

Appendix A

Using Proc REG

Model: MODEL1
Dependent Variable: VALUE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	74962789.378	74962789.378	20.510	0.0006
Error	13	47514727.556	3654979.0427		
C Total	14	122477516.93			

Root MSE	1911.79995	R-square	0.6121
Dep Mean	6596.26667	Adj R-sq	0.5822
C.V.	28.98306		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	3858.333333	780.48906065	4.943	0.0003
AUTO	1	4563.222222	1007.6070446	4.529	0.0006

Using Proc ANOVA

Analysis of Variance Procedure

Dependent Variable: VALUE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	74962789.37777770	74962789.37777770	20.51	0.0006
Error	13	47514727.55555550	3654979.04273505		
Corrected Total	14	122477516.93333300			

R-Square	C.V.	Root MSE	VALUE Mean
0.612053	28.98306	1911.79994841	6596.26666667

Source	DF	Anova SS	Mean Square	F Value	Pr > F
AUTO	1	74962789.37777780	74962789.37777780	20.51	0.0006

Using Proc ANOVA

General Linear Models Procedure

Dependent Variable: VALUE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	74962789.37777770	74962789.37777770	20.51	0.0006
Error	13	47514727.55555560	3654979.04273505		
Corrected Total	14	122477516.93333300			

R-Square	C.V.	Root MSE	VALUE Mean
0.612053	28.98306	1911.79994841	6596.26666667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AUTO	1	74962789.37777770	74962789.37777770	20.51	0.0006

Source	DF	Type III SS	Mean Square	F Value	Pr > F
AUTO	1	74962789.37777770	74962789.37777770	20.51	0.0006