Paper 264-26

Model Fitting in PROC GENMOD

Jean G. Orelien, Analytical Sciences, Inc.

Abstract:

There are several procedures in the SAS System for statistical modeling. Most statisticians who use the SAS system are familiar with procedures such as PROC REG and PROC GLM for fitting general linear models. However PROC GENMOD can handle these general linear models as well as more complex ones such as logistic models, loglinear models or models for count data. In addition, the main advantage of PROC GENMOD is that it can accommodate the analysis of correlated data. In this paper, we will discuss the use of PROC GENMOD to analyze simple as well as more complex statistical models. When other procedures are available to perform the same analysis, we will highlight the options from these procedures that may be missing in PROC GENMOD but might be of interest to the user. An example is given showing how PROC GENMOD is used to analyze various types of endpoints (continuous and count data) from a toxicology experiment. The materials in this paper should be accessible even to those users with limited data analysis skills.

1. Introduction

Generalized linear models (GLMs) include the most common statistical models used in Statistics. This class of models includes general linear models and logistics models. It should be noted that general linear models include ANOVA models as well as regression analysis. Complex models such as those arising from correlated data (repeated measures, clustered data) can also be fitted with GLMs. The form of a GLM model is given by:

$$f(Y)\ '\ X\beta\ \%\ \varepsilon \qquad (1)$$

The function *f* is known as the link distribution. For ANOVA, the link distribution is the identity. Other possible link functions include the logit for logistic regression or log for count data. Whereas in general linear models, it is necessary to assume that the errors are independent, have equal variances and are normally distributed, none of these assumptions are necessary in GLMs.

In the SAS System, GLMs can be fitted in PROC GENMOD. But separate procedures exist for certain sub classes of GLM models such as logistic regression or general linear models. For example, PROC GLM can fit general linear models. Although regression analysis can be fitted with PROC GLM, PROC REG is more specific to this type of analysis. Similarly, there exists also a PROC ANOVA that is specific as the name indicates to ANOVA models. Another SAS procedure for analyzing a subclass of GLM models is PROC LOGISTIC.

In this paper, we will provide an overview of some of the models that can be fitted with PROC GENMOD. When these models can be fitted by other SAS procedures, we will outline some differences between these procedures and GENMOD that the user needs to be aware of. In section 2, we discuss the fitting of GLM models in GENMOD and other procedures. The fitting of logistic models is discussed in section 3. Other types of models such as those involving other link distributions besides the logit and the identify as well as models for correlated data are discussed in section 4. In section 5, we provide an example showing how we have used PROC GENMOD to analyze different types of endpoints from the National Toxicology Program (NTP).

2. Fitting of General Linear Models in GENMOD and Other Procedures

There are many procedures besides PROC GENMOD in the SAS System for the fitting of general linear models. The three most commonly used are PROC ANOVA, PROC REG and PROC GLM. PROC GLM is the most comprehensive of the three models. Any analysis of general linear models can be performed in this procedure. However, the other two procedures are more efficient or offer more options for certain subclasses of general linear models. PROC ANOVA handles analysis of variance models with balanced designs. For these models, PROC ANOVA is faster than PROC GLM. For regression models, PROC REG provides many options that are not found in other procedures. Some of the advantages of PROC REG are that the user can fit several models with one call of the procedure, there are also more options for model selections and diagnostic tools to detect multicollinearity. Multicollinearity occurs when the independent variables in the models are correlated among themselves. This can lead to large variance for the estimated coefficients and affect our interpretation of these coefficients.

Although, PROC GENMOD can fit any general linear model, there are many useful options that it does not provide. For example, in an ANOVA model with random effects, one may be interested in estimating the variance components. This would not be possible in PROC GENMOD. With ANOVA models, one may also want to compare the means of an independent variable at several levels. As of version 7.0, there is an LSMEANS statement in PROC GENMOD, but unlike in PROC GLM, only the least squares estimates can be obtained with this statement, no statistical comparisons of the means can be made. It should be noted that while the other procedures that we have mentioned in this section used least squares methods to estimate the coefficient parameters, PROC GENMOD uses maximum likelihood methods. For general linear models, the maximum likelihood and the least squares methods yield the same estimates.

There are instances, where the data analyst is familiar with the data and may want to use PROC GENMOD to create outputs from the analysis of general linear models that are uniform with other outputs from more complex models that can only be analyzed in PROC GENMOD. An example of that will be given in section 5. In general, we suggest that GENMOD be used for analysis of GLM models only in those instances where the analyst wishes to obtain coefficient estimates and their variance. In these cases, the analyst should know from prior experience with similar data, that the data is well “behaved” and that no general linear models assumptions are violated.

We give below some examples of the use of PROC GENMOD for analysis of general linear models.

2.1 Example of a Regression analysis

Suppose sales for a company in a district can be predicted as a function of the target population and the per capita discretionary income. The data comes from Applied Linear Statistical Models from Neter, Wasserman and Kutner (page 249). Regression coefficients can be obtained with either of the following two syntaxes:

```
proc genmod data=salesdata;
model sales=target_population
discretionary_income;
```

or

```
proc reg data=salesdata;
model sales=target_population
discretionary_income;
```

2.2 Example of an Analysis of Variance

Suppose a research laboratory develops a new compound for the relief of severe hay fever and wants to compare the effect of the ingredients on the outcome. The outcome being measured is number hours of relief. This hypothetical example is also taken from Neter, Kutner and Wasserman (page 722). We could perform this analysis in PROC GENMOD with the following syntax:

```
proc genmod data=compounds_data;
class ingredient1 ingredient2;
model hours_of_relief=
ingredient1 ingredient2;
```

The same analysis could also be performed in PROC GLM:

```
proc glm data=compounds_data;
class ingredient1 ingredient2;
model hours_of_relief=
ingredient1 ingredient2;
```

3. Fitting of Logistic Models in PROC GENMOD and PROC LOGISTIC

Logistic models are of the form:

$$\log\left(\frac{p}{1\&p}\right) \,'\, \boldsymbol{X\beta} \;\%\; \varepsilon \qquad (2)$$

These models are appropriate for modeling proportions. Similar to a regular regression, a logistic model can be used to predict the proportion p that will be obtained for given values of the independent variables. But a logistic model can also be used to determine whether an independent variable significantly affects the variation of the dependent variable. In these cases, we are interested in knowing whether the odds of having the outcome are the same for all levels of the dependent variable(s). For example, we may be interested in determining if the odds of having a given disease is the same for smokers and nonsmokers. If one is interested in building a model to predict the variation of a proportion as a function of dependent variables, then PROC LOGISTIC would seem to be the clear choice because of the options it provides for model selections. On the other hand, if one is only interested in finding out whether an independent variable has a significant effect on the variation of a proportion then the analyst has the choice of using either PROC LOGISTIC or PROC GENMOD. In earlier versions of the SAS System [at least up to version 6.12], there was no CLASS or CONTRAST statements in PROC LOGISTIC. Thus, for the analysis of categorical variables one might have preferred PROC GENMOD over PROC LOGISTIC in earlier versions, since these categorical variables would have to be recoded in a data step prior to the call of the LOGISTIC procedure.

We give here an example of the use of PROC GENMOD for the analysis of binary data.

Example of a logistic regression analysis with binary data

Bliss (1935) reports the proportion of beetles killed after 5 hours of exposure at various concentrations of gaseous carbon disulphide. To obtain the regression coefficient to model proportion of beetles killed as a function of dosage, the following SAS code can be used:

```
proc genmod data=beetle_data;
model
number_killed/number_of_beetles=
dosage/link=logit dist=binomial;
```

PROC Logistic could also be used:

```
proc logistic data=beetle_data;
model
number_killed/number_of_beetles=
dosage/link=logit dist=binomial;
```

Notice that we needed to specify the LINK and DIST option since the defaults values used would not be appropriate for the analysis of binary data. The outcome under study is expressed as a ratio of two variables. Alternatively, we could use a dichotomous variable taking values 0 or 1 to indicate whether or not an individual beetle was killed and model that variable as a function of dosage of carbon disulphide.

4. Analysis of Count data and Correlated Data

The main advantage of PROC GENMOD compared to other data analysis procedures is the fact that it can fit complex models that cannot be fitted in other procedures for linear models such as GLM or Logistic. In this section, we discuss the use of proc GENMOD for the analysis of count data and correlated data. PROC GENMOD can fit data arising from a number of distributions. If the distribution is not available as an option, the user can even specify that distribution. One of the distributions available in PROC genmod is the Poisson distribution which is generally used for count data. Other available distributions include Gamma, Inverse Gaussian and Negative Binomial.

Correlated data can occur as the result of clustered data. Some examples of correlated data can occur as the result of taking repeated measurements on subjects or as a result of subjects belonging to the same cluster. The cluster can be a geographical region, a clinical site in a multi-site studies or a litter in a toxicity study. Failure to account for the correlation in the data can result in underestimating the variance which will lead to artificially low p-values. Several methods can be used to analyze correlated data including general linear multivariate models (GLMMs) or Linear Mixed Models. GLMM models can be fitted in PROC GLM and Linear mixed models can be fitted using PROC MIXED. Generalized estimating equations (GEE) methods which are used in GENMOD to account for correlated data in many situations may be preferred for various reasons (such as missing data or non-normality) over the other methods mentioned above.

For correlated data, the analyst must specify a "working correlation matrix". This working correlation matrix reflects the analyst assumption about the correlation structure between observations from the same cluster. The correlation structure can take many forms. One of the most commonly made assumptions is that the correlation within cluster is exchangeable. That is between any two elements of a cluster the correlation is the same. Other correlation structures that are available include: independent, autoregressive structure or m-dependent. The user can also specify a fixed correlation matrix.

The necessary information for Proc GENMOD to model the correlation in the data is inputted through the REPEATED statement. There are Options in the REPEATED statement to specify the form of the correlation structure as well as convergence criteria. The CORR option is probably the most important option in the REPEATED statement. It is used to specify the "working correlation matrix" that was described above and the SUBJECT option identifies the cluster.

Example

Paul (1982) reported an experiment in which pregnant rabbits were dosed with an unspecified toxic substance. The foetuses were observed for skeletal and visceral abnormalities. The cluster here is the litter. Typically, in these types of experiments, it is assumed that the correlation within litter is exchangeable, that is between any two littermates the correlation is the same. The data could be analyzed with the following syntax:

```
Proc genmod
data=rabbit_toxicity;
Class litter dose ;
Model malformation=dose /
link=logit ;
Repeated subject=litter/sorted
type=exch;
```

(The sorted option tells SAS that the data is properly sorted by subject.)

5. Example of the use of PROC GENMOD to analyze various types of endpoints from a toxicity study

In analyzing data from toxicity studies, my preference has been to use PROC GENMOD over other procedures. In this section, I will give a brief description of the data from these toxicology experiments. I will also discuss why we prefer to use PROC GENMOD over other procedures and how we use it.

A toxicology study can best be seen as a number of independent experiments of the effect of the same explanatory variable. For example to study the health effect of a given chemical, an experiment might be conducted to investigate the effect of this chemical on male reproductive organs and another experiment with the same chemical would investigate its effect on female reproductive organs. In most of these experiments such as the example from the previous section the data will be correlated and in some others, the data will be uncorrelated. For example in the reproductive toxicity studies conducted by the national institute of environmental health sciences (NIEHS) in some experiments, rodents selected from different litters are given the toxic substance and then are sacrificed. In these types of experiments, we would consider the data to be uncorrelated and traditional ANOVA methods could be used. The endpoints collected can be binary (such as malformation), continuous (such as organ weights) or count (such as sperm count). Even from the same experiment, there may be different types of endpoints.

In analyzing these data, we have preferred to use PROC GENMOD over other procedures. The main reason being that with the versatility of PROC GENMOD, we can handle correlated and uncorrelated data regardless of the type of endpoints. This makes it easier to analyze all of the endpoints from an experiment using a single SAS macro. For some of these endpoints the use of the PROC MIXED procedure might have been preferable. We opted against that option since doing so would have required that we use other procedures for endpoints that don't follow the normal distribution (binary and count data). Another reason for not using PROC MIXED is the fact that for some endpoints convergence can be difficult to achieve and would require a few trials and errors. Given that the number of endpoints to be analyzed can be more than 300, it is not possible to fit the endpoints one at a time.

To handle the large number of endpoints coming from these studies, we manipulated the data so that it could be sorted by endpoint, cluster and dose group. With the data sorted in this manner, the different type of endpoints (binary, continuous and count) were grouped together and analyzed in the same call to PROC GENMOD. For example, the continuous and count endpoints from an experiment where the data was correlated would be handled by the following SAS codes:

```
/* Correlated Data */

/* Continuous Endpoints */

proc genmod
data=toxdata(where=(count=0));
class dose litter;
model outcome=dose/type3
link=identity covb;
repeated subject=litter
/type=exch maxiter=25000 covb
corrb;
by endpt;

/* Count Endpoints */

proc genmod
data=toxdata(where=(count=1));
class dose litter;
model outcome=dose/type3
d=poisson covb;
repeated
subject=litter/type=exch
maxiter=25000 covb corrb;
```

```
by endpt;
```

In the SAS macro we have a macro variable to identify whether the observations from the experiments are correlated or not. For an experiment where the data were uncorrelated, the SAS code below would be executed by the macro:

```
/* Uncorrelated Data */

/* Continuous Endpoints */

proc genmod
data=toxdata(where=(count=0));
class dose;
model outcome=dose/type3
link=identity covb;
by endpt;

/* Count Endpoints */

proc genmod
data=toxdata(where=(count=1));
class dose litter;
model outcome=dose/type3
d=poisson covb;
by endpt;
```

Conclusion

Data that can be analyzed in PROC ANOVA, PROC GLM, PROC REG, or PROC LOGISTIC can also be handled in PROC GENMOD. There are instances where the analyst is familiar enough with the data and the only output of interest are the parameter estimates, standard errors , p-values and confidence intervals. In these instances, the use of PROC GENMOD might be preferred. We have given an example from toxicology data where using PROC GENMOD is more efficient because of the different type of endpoints that have to be analyzed from these experiments.

References

Agresti, A. Categorical Data Analysis (1990). New York: Wiley.

Neter J., Wasserman J. and Kutner M. (1990). Applied Linear Statistical Models. Boston: Irwin.

Orelien et al. (2000). Multiple Comparison with a control in GEE models using the SAS System. Presented at the SAS User Group International Conference in Indianapolis.

Stokes M.E., Davis C.S. and Koch G.G. (1995). Categorical Data Analysis Using the SAS System. Cary, SAS Institute, Inc.

Contact Information

Your questions and comments are welcome. Please contact:

Jean G. Orelien
Analytical Sciences, Inc.
2605 Meridian Pkwy.
Durham, NC 27713
Work Phone: (919)544-8500 (ext. 125)
Fax: (919)544-7307
Email: jorelien@asciences.com
Web: http://www.asciences.com