SAS® Software to Fit the Generalized Linear Model

Gordon Johnston, SAS Institute Inc., Cary, NC

Abstract

In recent years, the class of generalized linear models has gained popularity as a statistical modeling tool. This popularity is due in part to the flexibility of generalized linear models in addressing a variety of statistical problems and to the availability of software to fit the models. The SAS® system provides two new tools that fit generalized linear models. The GENMOD procedure in SAS/STAT® software is available in release 6.09 of the SAS system and in experimental form in release 6.08. SAS/INSIGHT® software provides a generalized linear modeling capability in release 6.08. This paper introduces generalized linear models and reviews the SAS software that fits the models.

Introduction

Generalized linear models are defined by Nelder and Wedderburn (1972). The class of generalized linear models is an extension of traditional linear models that allows the mean of a population to depend on a *linear predictor* through a nonlinear *link function* and allows the response probability distribution to be any member of an exponential family of distributions. Many widely used statistical models are generalized linear models. These include classical linear models with normal errors, logistic and probit models for binary data, and log-linear models for multinomial data. Many other useful statistical models can be formulated as generalized linear models by the selection of an appropriate link function and response probability distribution.

Refer to McCullagh and Nelder (1989) for a thorough account of statistical modeling using generalized linear models. The books by Aitkin, Anderson, Francis, and Hinde (1989) and Dobson (1990) are also excellent references with many examples of applications of generalized linear models. Firth (1991) provides an overview of generalized linear models.

What Is a Generalized Linear Model?

A traditional linear model is of the form

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i$$

where y_i is the response variable for the ith observation. The quantity $\mathbf{x}_i$ is a column vector of covariates, or explanatory variables for observation i, that is known from the experimental setting and is considered to be fixed, or non-random. The vector of unknown coefficients β is estimated by a least squares fit to the data $\mathbf{y}$. The ε_i are assumed to be independent, normal random variables with zero mean and constant variance. The expected value of y_i, denoted by μ_i, is

$$\mu_i = \mathbf{x}_i'\beta$$

While traditional linear models are used extensively in statistical data analysis, there are types of problems for which they are not appropriate.

- It may not be reasonable to assume that data are normally distributed. For example, the normal distribution (which is continuous) may not be adequate for modeling counts or measured proportions that are considered to be discrete.
- If the mean of the data is naturally restricted to a range of values, the traditional linear model may not be appropriate since the linear predictor $\mathbf{x}_i'\beta$ can take on any value. For example, the mean of a measured proportion is between 0 and 1, but the linear predictor of the mean in a traditional linear model is not restricted to this range.
- It may not be realistic to assume that the variance of the data is constant for all observations. For example, it is not unusual to observe data where the variance increases with the mean of the data.

A generalized linear model extends the traditional linear model and is therefore applicable to a wider range of data analysis problems. A generalized linear model consists of the following components.

- The linear component is defined just as it is for traditional linear models:

$$\eta_i = \mathbf{x}_i'\beta$$

- A monotonic differentiable link function g describes how the expected value of y_i is related to the linear predictor η_i:

$$g(\mu_i) = \mathbf{x}_i'\beta$$

- The response variables y_i are independent for $i = 1, 2, \ldots$ and have a probability distribution from an exponential family. This implies that the variance of the response depends on the mean μ through a *variance function* V:

$$var(y_i) = \phi V(\mu_i)/w_i$$

where ϕ is a constant and w_i is a known weight for each observation. The *dispersion parameter* ϕ is either known, for example for the binomial distribution, or it must be estimated.

As in the case of traditional linear models, fitted generalized linear models can be summarized through statistics such as parameter estimates, their standard errors, and goodness-of-fit statistics. You can also make statistical inference about the parameters using confidence intervals and hypothesis tests. However, specific inference procedures are usually based on asymptotic considerations, since exact distribution theory is not available or is not practical for all generalized linear models.

Examples of Generalized Linear Models

You construct a generalized linear model by deciding on response and explanatory variables for your data and choosing an appropriate link function and response probability distribution. Some examples of generalized linear models follow. Explanatory variables can be any combination of continuous variables, classification variables, and interactions.

Traditional Linear Model

- response variable: continuous variable
- distribution: normal
- link function: identity $\eta = \mu$

Logistic Regression

- response variable: a proportion
- distribution: binomial
- link function: logit $\eta = \log(\frac{\mu}{1-\mu})$

Poisson Regression in Log Linear Model

- response variable: a count
- distribution: Poisson
- link function: log $\eta = \log(\mu)$

Gamma Model with Log Link

- response variable: positive, continuous variable
- distribution: gamma
- link function: log $\eta = \log(\mu)$

The GENMOD Procedure

The GENMOD procedure fits a generalized linear model to the data by maximum likelihood estimation of the parameter vector β. There is, in general, no closed form solution for the maximum likelihood estimates of the parameters. The GENMOD procedure estimates the parameters of the model numerically through an iterative fitting process. The dispersion parameter ϕ is also estimated by maximum likelihood, or optionally by the residual deviance or by Pearson's chi-square divided by the degrees of freedom. Covariances, standard errors, and associated p-values are computed for the estimated parameters based on the asymptotic normality of maximum likelihood estimators.

A number of popular link functions and probability distributions are available in the GENMOD procedure. The built-in link functions are:

- identity: $\eta = \mu$
- logit: $\eta = \log(\mu/(1-\mu))$
- probit: $\eta = \Phi^{-1}(\mu)$, where Φ is the standard normal cumulative distribution function
- power: $\eta = \begin{cases} \mu^\lambda & \text{if } \lambda \neq 0 \\ \log(\mu) & \text{if } \lambda = 0 \end{cases}$
- log: $\eta = \log(\mu)$
- complementary log-log: $\eta = \log(-\log(1-\mu))$

The available distributions and associated variance functions are:

- normal: $V(\mu) = 1$
- binomial (proportion): $V(\mu) = \mu(1-\mu)$
- Poisson: $V(\mu) = \mu$
- gamma: $V(\mu) = \mu^2$
- inverse Gaussian: $V(\mu) = \mu^3$

In addition, you can easily define your own link functions or distributions through DATA step programming statements used within the procedure.

An important aspect of generalized linear modeling is the selection of explanatory variables in the model. Changes in goodness-of-fit statistics are often used to evaluate the contribution of subsets of explanatory variables to a particular model. The deviance, defined to be twice the difference between the maximum attainable log likelihood and the log likelihood of the model under consideration, is often used as a measure of goodness of fit. The maximum attainable log likelihood is achieved with a model that has a parameter for every observation.

One strategy for variable selection is to fit a sequence of models, beginning with a simple model with only an intercept term, and then include one additional explanatory variable in each successive model. You can measure the importance of the additional explanatory variable by the difference in deviances or fitted log likelihoods between successive models. Asymptotic tests computed by the GENMOD procedure allow you to assess the statistical significance of the additional term.

The GENMOD procedure allows you to fit a sequence of models, up through a maximum number of terms specified in a MODEL statement. A table summarizes likelihood ratio statistics for each successive pair of models. The likelihood ratio statistic for testing the significance of a subset of parameters in a model is defined as twice the difference in log likelihoods between the model and the submodel with the parameters set to zero. The asymptotic distribution of the likelihood ratio statistic is chi-square with degrees of freedom equal to the difference in the number of parameters between the model and submodel. *p*-values are computed in PROC GENMOD based on the asymptotic distributions of likelihood ratio statistics. This is called a *Type 1* analysis in the GENMOD procedure, because it is analogous to Type I (sequential) sums of squares in the GLM procedure. As with GLM Type I sums of squares, the results from this process depend on the order in which the model terms are fit.

The GENMOD procedure also generates a *Type 3* analysis analogous to Type III sums of squares in the GLM procedure. A Type 3 analysis does not depend on the order in which the terms for the model are specified. A GENMOD Type 3 analysis consists of specifying a model and computing likelihood ratio statistics for Type III contrasts for each term in the model. The contrasts are defined in the same way as they are in the GLM procedure. The GENMOD procedure optionally computes Wald statistics for Type III contrasts. This is computationally less expensive than likelihood ratio statistics, but it is thought to be less accurate because the specified significance level of hypothesis tests based on the Wald statistic may not be as close to the actual significance level as it is for likelihood ratio tests.

A Type 3 analysis generalizes the use of Type III estimable functions in linear models. Briefly, a Type III estimable function (contrast) for an effect is a linear function of the model parameters that involves the parameters of the effect and any interactions with that effect. A test of the hypothesis that the Type III contrast for a main effect is equal to 0 is intended to test the significance of the main effect in the presence of interactions. Refer to the documentation for the GLM procedure and Chapter 9, "The Four Types Of Estimable Functions," in *SAS/STAT User's Guide, Version 6, Fourth Edition* for more information about Type III estimable functions. Also, refer to *SAS System For Linear Models, Third Edition.*

Additional features of the GENMOD procedure are:

- likelihood ratio statistics for user-defined contrasts, that is, linear functions of the parameters, and *p*-values based on their asymptotic chi-square distributions
- ability to create a SAS data set corresponding to most tables printed by the procedure
- confidence intervals for model parameters based on either the profile likelihood function or asymptotic normality
- PROC GLM-like syntax for the specification of the response and model effects, including interaction terms and automatic coding of classification variables

Poisson Regression

You can use the GENMOD procedure to fit a variety of statistical models. A typical use of the GENMOD procedure is to perform Poisson regression.

The Poisson distribution can be used to model the distribution of cell counts in a multiway contingency table. Aitkin, Anderson, Francis, and Hinde (1989) have used this method to model insurance claims data. Suppose the following hypothetical insurance claims data are classified by two factors: age group, with two levels, and car type, with three levels.

```
data insure;
   input n c car$ age;
   ln = log(n);
   cards;
   500   42  small  1
   1200  37  medium 1
   100    1  large  1
   400  101  small  2
   500   73  medium 2
   300   14  large  2
   ;
```

In the preceding data set, N is the number of insurance policyholders, and C is the number of insurance claims. CAR is the type of car involved, classified into three groups, and AGE is the age group of a policyholder, classified into two groups.

You can use the GENMOD procedure to perform a Poisson regression analysis of these data with a log link function. Assume the number of claims C has a Poisson probability distribution, and its mean, μ_i, is related to the factors CAR and AGE for observation i by

$$\begin{aligned}\log(\mu_i) &= \log(N_i) + \beta_0 + CAR_i(1)\beta_1 + \\ &\quad CAR_i(2)\beta_2 + CAR_i(3)\beta_3 + \\ &\quad AGE_i(1)\beta_4 + AGE_i(2)\beta_5\end{aligned}$$

$CAR_i(j)$ and $AGE_i(j)$ are indicator variables associated with the jth level of CAR and AGE:

$$CAR_i(j) = \begin{cases} 1 & \text{if CAR} = j \\ 0 & \text{if CAR} \neq j \end{cases}$$

for observation i. The βs are unknown parameters to be estimated by the procedure. The logarithm of N is used as an *offset*, that is, a regression variable with a constant coefficient of 1 for each observation. A log linear relationship between the mean and the factors CAR and AGE is specified by the log link function. The log link function insures that the mean number of insurance claims for each car and age group predicted from the fitted model will be positive.

The following statements invoke the GENMOD procedure to perform this analysis.

```
proc genmod data=insure;
   class car age;
   model c = car age / dist   = poisson
                       link   = log
                       offset = ln;
   run;
```

CAR and AGE are specified as CLASS variables so that PROC GENMOD automatically generates the indicator variables associated with CAR and AGE.

The MODEL statement specifies C as the response variable and CAR and AGE as explanatory variables. An intercept term is included by default. Thus, the model matrix **X** (the matrix that has as its ith row the transpose of the covariate vector for the ith observation) consists of a column of 1s representing the intercept term and columns of 0s and 1s derived from indicator variables representing the levels of the CAR and AGE variables. That is, the model matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

where the first column corresponds to the intercept, the next 3 columns correspond to CAR, and the last 2 columns correspond to AGE.

The response distribution is specified as Poisson, and the link function is chosen to be log. That is, the Poisson mean parameter μ is related to the linear predictor by

$$\log(\mu) = \mathbf{x}_i'\boldsymbol{\beta}.$$

The logarithm of N is specified as an offset variable, as is common in this type of analysis. In this case the offset variable serves to normalize the fitted cell means to a per policyholder basis, since the total number of claims, not individual policyholder claims, were observed.

PROC GENMOD produces the following default output from the preceding statements.

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.INSURE
Distribution	POISSON
Link Function	LOG
Dependent Variable	C
Offset Variable	LN
Observations Used	6

Figure 1. Model Information

The "Model Information" table in Figure 1 provides information about the specified model and the input data set.

```
The GENMOD Procedure

Class Level Information

Class      Levels  Values

CAR             3  large medium small
AGE             2  1 2
```

Figure 2. Class Level Information

The "Class Level Information" table in Figure 2 identifies the levels of the classification variables that are used in the model. Note that CAR is a character variable, and the values are sorted in alphabetical order. This is the default sort order, but you can select different sort orders with the ORDER= option in the PROC GENMOD statement.

```
The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion              DF          Value

Deviance                2          2.8207
Scaled Deviance         2          2.8207
Pearson Chi-Square      2          2.8416
Scaled Pearson X2       2          2.8416
Log Likelihood          .        837.4533


Value/DF

  1.4103
  1.4103
  1.4208
  1.4208
```

Figure 3. Goodness of Fit Criteria

The "Criteria For Assessing Goodness Of Fit" table in Figure 3 contains statistics that summarize the fit of the specified model. These statistics are helpful in judging the adequacy of a model and in comparing it with other models under consideration. If you compare the deviance of 2.8207 with its asymptotic chi-square with 2 degrees of freedom distribution, you find that the *p*-value is .24. This indicates that the specified model fits the data reasonably well.

```
The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter              DF    Estimate

INTERCEPT               1     -1.3168
CAR         large       1     -1.7643
CAR         medium      1     -0.6928
CAR         small       0      0.0000
AGE         1           1     -1.3199
AGE         2           0      0.0000
SCALE                   0      1.0000
NOTE:  The scale parameter was held fixed.

  Std Err   ChiSquare  Pr>Chi

   0.0903    212.7321  0.0000
   0.2724     41.9587  0.0000
   0.1282     29.1800  0.0000
   0.0000           .       .
   0.1359     94.3388  0.0000
   0.0000           .       .
   0.0000           .       .
```

Figure 4. Parameter Estimates

The "Analysis Of Parameter Estimates" table in Figure 4 summarizes the results of the iterative parameter estimation process. For each parameter in the model, PROC GENMOD prints columns with the parameter name, the degrees of freedom associated with the parameter, the estimated parameter value, the standard error of the parameter estimate, and a Wald chi-square statistic and associated *p*-value for testing the significance of the parameter to the model. If a column of the model matrix corresponding to a parameter is found to be linearly dependent, or *aliased*, with columns corresponding to parameters preceding it in the model, PROC GENMOD assigns it zero degrees of freedom and prints a value of zero for both the parameter estimate and its standard error.

This table includes a row for a scale parameter, even though there is no free scale parameter in the Poisson distribution. PROC GENMOD allows the specification of a scale parameter to fit overdispersed Poisson and binomial distributions. In such cases, the SCALE row indicates the value of the overdispersion scale parameter used in adjusting output statistics. PROC GENMOD prints a note indicating that the scale parameter was fixed, that is, not estimated by the iterative fitting process.

It is usually of interest to assess the importance of the main effects in the model. Type 1 and Type 3 analyses generate statistical tests for the significance of these effects. You can request these analyses with the TYPE1 and TYPE3 options in the MODEL statement.

```
proc genmod data=insure;
   class car age;
   model c = car age / dist   = poisson
                       link   = log
                       offset = ln
                       type1
                       type3;
   run;
```

The results of these analyses are summarized in the tables that follow.

The GENMOD Procedure

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	175.1536	0	.	.
CAR	107.4620	2	67.6915	0.0000
AGE	2.8207	1	104.6414	0.0000

Figure 5. Type 1 Analysis

In the table for Type 1 analysis in Figure 5, each entry in the deviance column represents the deviance for the model containing the effect for that row and all effects preceding it in the table. For example, the deviance corresponding to CAR in the table is the deviance of the model containing an intercept and CAR. As more terms are included in the model, the deviance decreases.

Entries in the chi-square column are likelihood ratio statistics for testing the significance of the effect added to the model containing all the preceding effects. The chi-square value of 67.6915 for CAR represents twice the difference in log likelihoods between fitting a model with only an intercept term and a model with an intercept and CAR. Since the scale parameter is set to 1 in this analysis, this is equal to the difference in deviances. Since two additional parameters are involved, this statistic can be compared with a chi-square distribution with two degrees of freedom. The resulting *p*-value (labeled Pr>Chi) of 0 indicates that this variable is highly significant. Similarly, the chi-square value of 104.6414 for AGE represents the difference in log likelihoods between the model with the intercept and CAR and the model with the intercept, CAR, and AGE. This effect is also highly significant, as indicated by the *p*-value.

The GENMOD Procedure

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
CAR	2	72.8181	0.0000
AGE	1	104.6414	0.0000

Figure 6. Type 3 Analysis

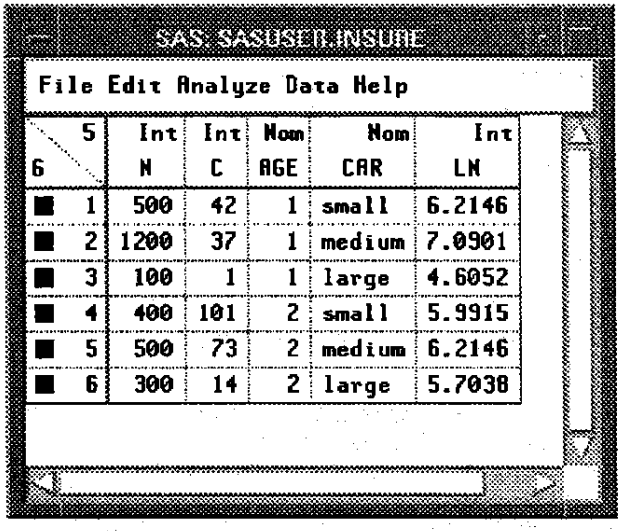
The Type 3 analysis shown in Figure 6 results in the same conclusions as the Type 1 analysis. The Type 3 chi-square value for CAR, for example, is twice the difference between the log likelihood for the model with INTERCEPT, CAR, and AGE included and the log likelihood for the model with CAR excluded. The hypothesis tested in this case is the significance of CAR in the model with AGE already included.

The values of the Type 3 likelihood ratio statistics for CAR and AGE indicate that both of these factors are highly significant in determining the claims performance of the insurance policyholders.

SAS/INSIGHT Software

You can fit generalized linear models within an interactive graphical environment using SAS/INSIGHT software. The same set of response distributions and link functions, with the exception of user-defined, are available in SAS/INSIGHT software as in the GENMOD procedure. Most of the output statistics in PROC GENMOD are also available in SAS/INSIGHT software, and some additional regression diagnostics and automatic plotting of residuals are available.

The SAS/INSIGHT data window containing the insurance claims data is shown in Figure 7. CAR and AGE have been selected as nominal, or CLASS variables.



6 \ 5	Int N	Int C	Nom AGE	Nom CAR	Int LN
1	500	42	1	small	6.2146
2	1200	37	1	medium	7.0901
3	100	1	1	large	4.6052
4	400	101	2	small	5.9915
5	500	73	2	medium	6.2146
6	300	14	2	large	5.7038

Figure 7. Data Window

You select **Analyze → Fit(Y X)** to invoke the window shown in Figure 8. There you can select the response variable and covariate variables by selecting the variable names and then clicking the **Y** button for the response variable and the **X** button for the covariates. C has been selected as the response, and CAR and AGE have been selected as covariates.

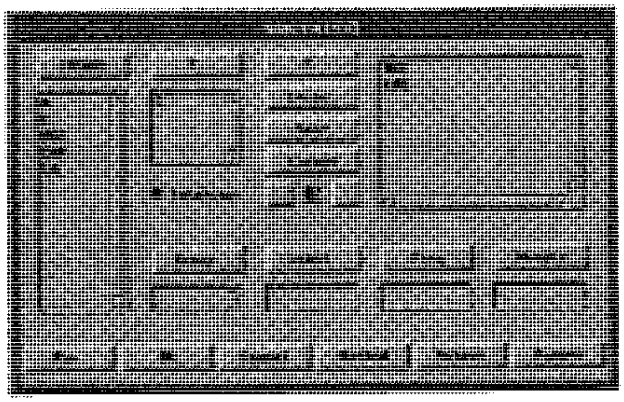


Figure 8. Specifying the Model

You can then click the **Method** button to specify the generalized linear model in the window shown in Figure 9. The Poisson response distribution and log link function have been selected. You specify LN as an offset variable by selecting the variable name and then clicking the **Offset** button.

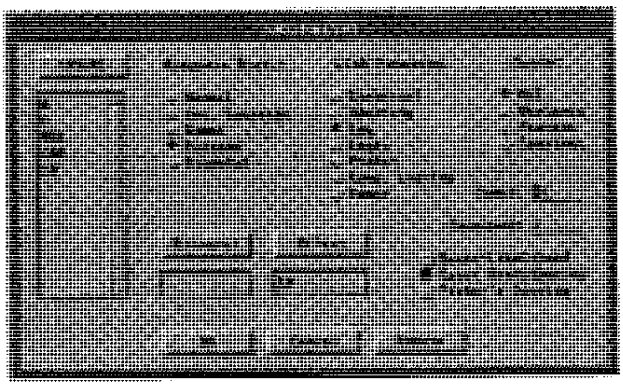


Figure 9. Selecting the Response Distribution and Link Function

You can select the output you desire from the analysis by clicking the **Output** button in Figure 8. This invokes the output window shown in Figure 10. As shown in Figure 10, Type I tests, Type III tests, and Parameter Estimates have been selected.

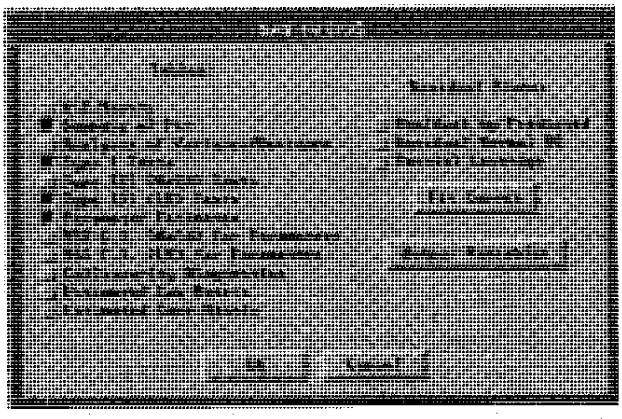


Figure 10. Selecting the Output

The results are shown in the analysis output window in Figure 11. These are identical to the results produced by PROC GENMOD.

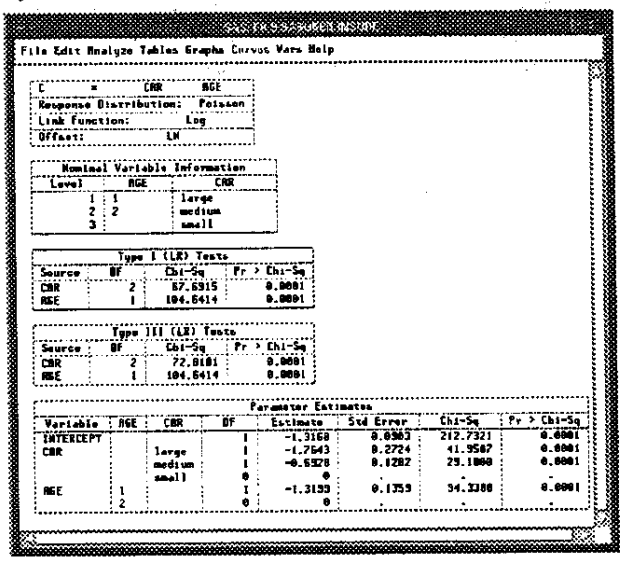


Figure 11. Analysis Results

Conclusions

The generalized linear model extends the traditional linear model to be applicable to a wider range of statistical modeling problems. The GENMOD procedure in SAS/STAT software fits generalized linear models in a traditional SAS environment, retaining much of the syntax and functionality of linear modeling procedures such as PROC GLM. You can also fit generalized linear models in an interactive graphical interface environment using SAS/INSIGHT software. Both methods produce statistics that allow you to make statistical inference about the model parame-

ters.

References

Aitkin, M., Anderson, D., Francis, B., and Hinde, J. (1989), *Statistical Modelling in GLIM*, Oxford: Oxford Science Publications.

Dobson, A. (1990), *An Introduction To Generalized Linear Models*, London: Chapman and Hall.

Firth, D. (1991), "Generalized Linear Models," in *Statistical Theory and Modelling*, ed. Hinkley, D.V., Reid, N., and Snell, E.J., London: Chapman and Hall.

McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, London: Chapman and Hall.

Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society A*, 135, 370-384.