Chapter 29
# The GENMOD Procedure

## Chapter Table of Contents

# Chapter 29
# The GENMOD Procedure

## Overview

The GENMOD procedure fits generalized linear models, as 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 discussion 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.

The analysis of correlated data arising from repeated measurements when the measurements are assumed to be multivariate normal has been studied extensively. However, the normality assumption may not always be reasonable; for example, different methodology must be used in the data analysis when the responses are discrete and correlated. Generalized Estimating Equations (GEEs) provide a practical method with reasonable statistical efficiency to analyze such data.

Liang and Zeger (1986) introduced GEEs as a method of dealing with correlated data when, except for the correlation among responses, the data can be modeled as a generalized linear model. For example, correlated binary and count data in many cases can be modeled in this way.

The GENMOD procedure can fit models to correlated responses by the GEE method. You can use PROC GENMOD to fit models with most of the correlation structures from Liang and Zeger (1986) using GEEs. Refer to Liang and Zeger (1986), Diggle, Liang, and Zeger (1994), and Lipsitz, Fitzmaurice, Orav, and Laird (1994) for more details on GEEs.

## What is a Generalized Linear Model?

A traditional linear model is of the form

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

where $y_i$ is the response variable for the $i$th 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 nonrandom. The vector of unknown coefficients $\boldsymbol{\beta}$ is estimated by a least squares fit to the data $\mathbf{y}$. The $\varepsilon_i$ are assumed to be independent, normal random variables with zero mean and constant variance. The expected value of $y_i$, denoted by $\mu_i$, is

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

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 $\mathbf{x}_i{}'\boldsymbol{\beta}$ 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:

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

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

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

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

  $$var(y_i) = \frac{\phi V(\mu_i)}{w_i}$$

where $\phi$ is a constant and $w_i$ is a known weight for each observation. The *dispersion parameter* $\phi$ is either known (for example, for the binomial or Poisson distribution, $\phi = 1$) or it must be estimated.

See the section "Response Probability Distributions" on page 1402 for the form of a probability distribution from the exponential family of distributions.

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: a continuous variable
- distribution: normal
- link function: identity $\quad g(\mu) = \mu$

### Logistic Regression

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

### Poisson Regression in Log Linear Model

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

### Gamma Model with Log Link

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

# The GENMOD Procedure

The GENMOD procedure fits a generalized linear model to the data by maximum likelihood estimation of the parameter vector $\boldsymbol{\beta}$. 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 $\phi$ 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 $p$-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: $g(\mu) = \mu$
- logit: $g(\mu) = \log(\mu/(1-\mu))$
- probit: $g(\mu) = \Phi^{-1}(\mu)$, where $\Phi$ is the standard normal cumulative distribution function
- power: $g(\mu) = \begin{cases} \mu^{\lambda} & \text{if } \lambda \neq 0 \\ \log(\mu) & \text{if } \lambda = 0 \end{cases}$
- log: $g(\mu) = \log(\mu)$
- complementary log-log: $g(\mu) = \log(-\log(1-\mu))$

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: $V(\mu) = \mu^3$
- negative binomial: $V(\mu) = \mu + k\mu^2$
- multinomial

The negative binomial is a distribution with an additional parameter $k$ in the variance function. PROC GENMOD estimates $k$ by maximum likelihood, or you can optionally set it to a constant value. Refer to McCullagh and Nelder (1989, Chapter 11), Hilbe (1994), or Lawless (1987) for discussions of the negative binomial distribution.

The multinomial distribution is sometimes used to model a response that can take values from a number of categories. The binomial is a special case of the multinomial with two categories. See the section "Multinomial Models" on page 1419 and refer to McCullagh and Nelder (1989, Chapter 5) for a description of the multinomial distribution.

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. See the section "Goodness of Fit" on page 1408 for formulas for the deviance.

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 enable you to assess the statistical significance of the additional term.

The GENMOD procedure enables you to fit a sequence of models, up through a maximum number of terms specified in a MODEL statement. A table summarizes twice the difference in log likelihoods between each successive pair of models. 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 the PROC 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 procedure 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. See Chapter 30, "The GLM Procedure," and Chapter 12, "The Four Types of Estimable Functions," for more information about Type III estimable functions. Also refer to Littell, Freund, and Spector (1991).

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

- estimated values, standard errors, and confidence limits for user-defined contrasts and least-squares means
- ability to create a SAS data set corresponding to most tables displayed by the procedure (see Table 29.3 on page 1438)
- confidence intervals for model parameters based on either the profile likelihood function or asymptotic normality
- syntax similar to that of PROC GLM for the specification of the response and model effects, including interaction terms and automatic coding of classification variables
- ability to fit GEE models for clustered response data

# Getting Started

## Poisson Regression

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

You can use the Poisson distribution 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);
   datalines;
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, the variable n represents the number of insurance policyholders and the variable c represents the number of insurance claims. The variable car is the type of car involved (classified into three groups) and the variable age is the age group of a policyholder (classified into two groups).

You can use PROC GENMOD to perform a Poisson regression analysis of these data with a log link function. This type of model is sometimes called a *loglinear model*.

Assume that the number of claims c has a Poisson probability distribution and that its mean, $\mu_i$, is related to the factors car and age for observation $i$ by

$$\begin{aligned}\log(\mu_i) &= \log(\mathrm{n}_i) + \mathbf{x}_i'\boldsymbol{\beta} \\ &= \log(\mathrm{n}_i) + \beta_0 + \\ &\quad \mathrm{car}_i(1)\beta_1 + \mathrm{car}_i(2)\beta_2 + \mathrm{car}_i(3)\beta_3 + \\ &\quad \mathrm{age}_i(1)\beta_4 + \mathrm{age}_i(2)\beta_5\end{aligned}$$

The indicator variables $\mathrm{car}_i(j)$ and $\mathrm{age}_i(j)$ are associated with the $j$th level of the variables car and age for observation $i$

$$\mathrm{car}_i(j) = \begin{cases} 1 & \text{if car} = j \\ 0 & \text{if car} \neq j \end{cases}$$

The $\beta$s are unknown parameters to be estimated by the procedure. The logarithm of the variable 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 ensures that the mean number of insurance claims for each car and age group predicted from the fitted model is 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;
```

The variables 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 $\mathbf{X}$ (the matrix which has as its $i$th row the transpose of the covariate vector for the $i$th 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

$$\mathbf{X} = \left[\begin{array}{c|ccc|cc} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{array}\right]$$

where the first column corresponds to the intercept, the next three columns correspond to the variable car, and the last two columns correspond to the variable age.

The response distribution is specified as Poisson, and the link function is chosen to be log. That is, the Poisson mean parameter $\mu$ 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, are observed.

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

```
                          The SAS System

                       The GENMOD Procedure

                        Model Information

                Data Set              WORK.INSURE
                Distribution              Poisson
                Link Function                 Log
                Dependent Variable              c
                Offset Variable                ln
                Observations Used               6
```

**Figure 29.1.** Model Information

The "Model Information" table displayed in Figure 29.1 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 29.2.** Class Level Information

Figure 29.2 displays the "Class Level Information" table, which identifies the levels of the classification variables that are used in the model. Note that car is a charac-

ter 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 (see the ORDER= option on page 1379 for details).

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 2 | 2.8207 | 1.4103 |
| Scaled Deviance | 2 | 2.8207 | 1.4103 |
| Pearson Chi-Square | 2 | 2.8416 | 1.4208 |
| Scaled Pearson X2 | 2 | 2.8416 | 1.4208 |
| Log Likelihood | | 837.4533 | |

**Figure 29.3.** Goodness Of Fit

The "Criteria For Assessing Goodness Of Fit" table displayed in Figure 29.3 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 0.24. This indicates that the specified model fits the data reasonably well.

The GENMOD Procedure

Analysis Of Parameter Estimates

| Parameter | | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square |
|---|---|---|---|---|---|---|---|
| Intercept | | 1 | -1.3168 | 0.0903 | -1.4937 | -1.1398 | 212.73 |
| car | large | 1 | -1.7643 | 0.2724 | -2.2981 | -1.2304 | 41.96 |
| car | medium | 1 | -0.6928 | 0.1282 | -0.9441 | -0.4414 | 29.18 |
| car | small | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| age | 1 | 1 | -1.3199 | 0.1359 | -1.5863 | -1.0536 | 94.34 |
| age | 2 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Scale | | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | |

Analysis Of Parameter Estimates

| Parameter | | Pr > ChiSq |
|---|---|---|
| Intercept | | <.0001 |
| car | large | <.0001 |
| car | medium | <.0001 |
| car | small | . |
| age | 1 | <.0001 |
| age | 2 | . |
| Scale | | |

NOTE: The scale parameter was held fixed.

**Figure 29.4.** Analysis Of Parameter Estimates

Figure 29.4 displays the "Analysis Of Parameter Estimates" table, which summarizes the results of the iterative parameter estimation process. For each parameter in the model, PROC GENMOD displays columns with the parameter name, the degrees of freedom associated with the parameter, the estimated parameter value, the standard

error of the parameter estimate, the confidence intervals, and the 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 displays 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. See the "Response Probability Distributions" section on page 1402 for the form of the Poisson probability 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. See the section "Overdispersion" on page 1410 for more on overdispersion and the meaning of the SCALE parameter output by the GENMOD procedure. PROC GENMOD displays a note indicating that the scale parameter is 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 | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Intercept | 175.1536 | | | |
| car | 107.4620 | 2 | 67.69 | <.0001 |
| age | 2.8207 | 1 | 104.64 | <.0001 |

**Figure 29.5.** Type 1 Analysis

In the table for Type 1 analysis displayed in Figure 29.5, 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 67.69 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 less than 0.0001 indicates that this variable is highly significant. Similarly, the chi-square value of 104.64 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 small $p$-value.

**The GENMOD Procedure**

**LR Statistics For Type 3 Analysis**

| Source | DF | Chi-Square | Pr > ChiSq |
|---|---|---|---|
| car | 2 | 72.82 | <.0001 |
| age | 1 | 104.64 | <.0001 |

**Figure 29.6.** Type 3 Analysis

The Type 3 analysis results in the same conclusions as the Type 1 analysis. The Type 3 chi-square value for the car variable, for example, is twice the difference between the log likelihood for the model with the variables Intercept, car, and age included and the log likelihood for the model with the car variable excluded. The hypothesis tested in this case is the significance of the variable car given that the variable age is in the model. In other words, it tests the additional contribution of car in the model.

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

## Generalized Estimating Equations

This section illustrates the use of the REPEATED statement to fit a GEE model, using repeated measures data from the "Six Cities" study of the health effects of air pollution (Ware et al. 1984). The data analyzed are the 16 selected cases in Lipsitz, Fitzmaurice, et al. (1994). The binary response is the wheezing status of 16 children at ages 9, 10, 11, and 12 years. The mean response is modeled as a logistic regression model using the explanatory variables city of residence, age, and maternal smoking status at the particular age. The binary responses for individual children are assumed to be equally correlated, implying an exchangeable correlation structure.

The data set and SAS statements that fit the model by the GEE method are as follows:

```
data six;
   input case city$ @;
   do i=1 to 4;
      input age smoke wheeze @@;
      output;
   end;
   datalines;
 1 portage   9 0 1  10 0 1  11 0 1  12 0 0
 2 kingston  9 1 1  10 2 1  11 2 0  12 2 0
 3 kingston  9 0 1  10 0 0  11 1 0  12 1 0
 4 portage   9 0 0  10 0 1  11 0 1  12 1 0
 5 kingston  9 0 0  10 1 0  11 1 0  12 1 0
 6 portage   9 0 0  10 1 0  11 1 0  12 1 0
 7 kingston  9 1 0  10 1 0  11 0 0  12 0 0
 8 portage   9 1 0  10 1 0  11 1 0  12 2 0
 9 portage   9 2 1  10 2 0  11 1 0  12 1 0
10 kingston  9 0 0  10 0 0  11 0 0  12 1 0
11 kingston  9 1 1  10 0 0  11 0 1  12 0 1
12 portage   9 1 0  10 0 0  11 0 0  12 0 0
13 kingston  9 1 0  10 0 1  11 1 1  12 1 1
14 portage   9 1 0  10 2 0  11 1 0  12 2 1
15 kingston  9 1 0  10 1 0  11 1 0  12 2 1
16 portage   9 1 1  10 1 1  11 2 0  12 1 0
;
proc genmod data=six ;
   class case city ;
   model  wheeze = city age smoke  /  dist=bin;
   repeated  subject=case / type=exch covb corrw;
run;
```

The CLASS statement and the MODEL statement specify the model for the mean of the wheeze variable response as a logistic regression with city, age, and smoke as independent variables, just as for an ordinary logistic regression.

The REPEATED statement invokes the GEE method, specifies the correlation structure, and controls the displayed output from the GEE model. The option SUBJECT=CASE specifies that individual subjects are identified in the input data set by the variable case. The SUBJECT= variable case must be listed in the CLASS statement. Measurements on individual subjects at ages 9, 10, 11, and 12 are in the proper order in the data set, so the WITHINSUBJECT= option is not required. The TYPE=EXCH option specifies an exchangeable working correlation structure, the COVB option specifies that the parameter estimate covariance matrix be displayed, and the CORRW option specifies that the final working correlation be displayed.

Initial parameter estimates for iterative fitting of the GEE model are computed as in an ordinary generalized linear model, as described previously. Results of the initial model fit displayed as part of the generated output are not shown here. Statistics for the initial model fit such as parameter estimates, standard errors, deviances, and Pearson chi-squares do not apply to the GEE model, and are only valid for the initial model fit. The following tables display information that applies to the GEE model fit.

Figure 29.7 displays general information about the GEE model fit.

**The GENMOD Procedure**

**GEE Model Information**

| | |
|---|---|
| Correlation Structure | Exchangeable |
| Subject Effect | case (16 levels) |
| Number of Clusters | 16 |
| Correlation Matrix Dimension | 4 |
| Maximum Cluster Size | 4 |
| Minimum Cluster Size | 4 |

**Figure 29.7.** GEE Model Information

Figure 29.8 displays the parameter estimate covariance matrices specified by the COVB option. Both model-based and empirical covariances are produced.

**The GENMOD Procedure**

**Covariance Matrix (Model-Based)**

| | Prm1 | Prm2 | Prm4 | Prm5 |
|---|---|---|---|---|
| Prm1 | 5.74947 | -0.22257 | -0.53472 | 0.01655 |
| Prm2 | -0.22257 | 0.45478 | -0.002410 | 0.01876 |
| Prm4 | -0.53472 | -0.002410 | 0.05300 | -0.01658 |
| Prm5 | 0.01655 | 0.01876 | -0.01658 | 0.19104 |

**Covariance Matrix (Empirical)**

| | Prm1 | Prm2 | Prm4 | Prm5 |
|---|---|---|---|---|
| Prm1 | 9.33994 | -0.85104 | -0.83253 | -0.16534 |
| Prm2 | -0.85104 | 0.47368 | 0.05736 | 0.04023 |
| Prm4 | -0.83253 | 0.05736 | 0.07778 | -0.002364 |
| Prm5 | -0.16534 | 0.04023 | -0.002364 | 0.13051 |

**Figure 29.8.** GEE Parameter Estimate Covariance Matrices

The exchangeable working correlation matrix specified by the CORRW option is displayed in Figure 29.9.

**The GENMOD Procedure**

**Working Correlation Matrix**

| | Col1 | Col2 | Col3 | Col4 |
|---|---|---|---|---|
| Row1 | 1.0000 | 0.1648 | 0.1648 | 0.1648 |
| Row2 | 0.1648 | 1.0000 | 0.1648 | 0.1648 |
| Row3 | 0.1648 | 0.1648 | 1.0000 | 0.1648 |
| Row4 | 0.1648 | 0.1648 | 0.1648 | 1.0000 |

**Figure 29.9.** GEE Working Correlation Matrix

The parameter estimates table, displayed in Figure 29.10, contains parameter estimates, standard errors, confidence intervals, $Z$ scores, and $p$-values for the parameter estimates. Empirical standard error estimates are used in this table. A table using model-based standard errors can be created by using the REPEATED statement option MODELSE.

The GENMOD Procedure

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

| Parameter | | Estimate | Standard Error | 95% Confidence Limits | | Z | Pr > \|Z\| |
|---|---|---|---|---|---|---|---|
| Intercept | | 1.2751 | 3.0561 | -4.7148 | 7.2650 | 0.42 | 0.6765 |
| city | kingston | 0.1223 | 0.6882 | -1.2266 | 1.4713 | 0.18 | 0.8589 |
| city | portage | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| age | | -0.2036 | 0.2789 | -0.7502 | 0.3431 | -0.73 | 0.4655 |
| smoke | | -0.0935 | 0.3613 | -0.8016 | 0.6145 | -0.26 | 0.7957 |

**Figure 29.10.** GEE Parameter Estimates Table

# Syntax

You can specify the following statements in the GENMOD procedure. Items within the <> are optional.

**PROC GENMOD** < *options* > **;**
**BY** *variables* **;**
**CLASS** *variables* **;**
**CONTRAST** *'label' effect values* < *... effect values* > < */options* > **;**
**DEVIANCE** *variable = expression* **;**
**ESTIMATE** *'label' effect values* < *... effect values* > < */options* > **;**
**FREQ** | **FREQUENCY** *variable* **;**
**FWDLINK** *variable = expression* **;**
**INVLINK** *variable = expression* **;**
**LSMEANS** *effects* < */ options* > **;**
**MAKE** *'table'* **OUT=***SAS-data-set***;**
**OUTPUT** < **OUT=***SAS-data-set* >
< *keyword=name...keyword=name* > **;**
**MODEL** *response* = < *effects* >< */options* > **;**
*programming statements*
**REPEATED SUBJECT=** *subject-effect* < */ options* > **;**
**WEIGHT** | **SCWGT** *variable* **;**
**VARIANCE** *variable = expression* **;**

The PROC GENMOD statement invokes the procedure. All statements other than the MODEL statement are optional. The CLASS statement, if present, must precede the MODEL statement, and the CONTRAST statement must come after the MODEL statement.

## PROC GENMOD Statement

**PROC GENMOD**< *options* > **;**

The PROC GENMOD statement invokes the procedure. You can specify the following options.

**DATA=***SAS-data-set*
specifies the SAS data set containing the data to be analyzed. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

**DESCENDING | DESC**
specifies that the levels of the response variable for the ordinal multinomial model be sorted in the reverse of the default order. For example, if RORDER=FORMATTED (the default), the DESCENDING option causes the levels to be sorted from highest to lowest instead of from lowest to highest. If RORDER=FREQ, the DESCENDING option causes the levels to be sorted from lowest frequency count to highest instead of from highest to lowest.

**NAMELEN=**$n$
specifies the length of effect names in tables and output data sets to be $n$ characters long, where $n$ is a value between 20 and 200 characters. The default length is 20 characters.

**ORDER=***keyword*
specifies the sorting order for the levels of the classification variables (specified in the CLASS statement). This ordering determines which parameters in the model correspond to each level in the data, so the ORDER= option may be useful when you use the CONTRAST or ESTIMATE statement. Note that the ORDER= option applies to the levels for all classification variables. The exception is ORDER=FORMATTED (the default) for numeric variables for which you have supplied no explicit format (that is, for which there is no corresponding FORMAT statement in the current PROC GENMOD run or in the DATA step that created the data set). In this case, the levels are ordered by their internal (numeric) value. Note that this represents a change from previous releases for how class levels are ordered. In releases previous to Version 8, numeric class levels with no explicit format were ordered by their BEST12. formatted values, and in order to revert to the previous ordering you can specify this format explicitly for the affected classification variables. The change was implemented because the former default behavior for ORDER=FORMATTED often resulted in levels not being ordered numerically and usually required the user to intervene with an explicit format or ORDER=INTERNAL to get the more natural ordering. The following table displays the valid *keywords* and describes how PROC GENMOD interprets them.

| **ORDER=** *keyword* | **Levels Sorted by** |
|---|---|
| DATA | order of appearance in the input data set |
| FORMATTED | external formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value |
| FREQ | descending frequency count; levels with the most observations come first in the order |
| INTERNAL | unformatted value |

By default, ORDER=FORMATTED. For ORDER=FORMATTED and ORDER=INTERNAL, the sort order is machine dependent. For more information on sorting order, refer to the chapter titled "The SORT Procedure" in the *SAS Procedures Guide*.

**RORDER=***keyword*

specifies the sorting order for the levels of the response variable. This ordering determines which intercept parameter in the model corresponds to each level in the data. If RORDER=FORMATTED (the default) for numeric variables for which you have supplied no explicit format (that is, for which there is no corresponding FORMAT statement in the current PROC GENMOD run or in the DATA step that created the data set), the levels are ordered by their internal (numeric) value. Note that this represents a change from previous releases for how class levels are ordered. In releases previous to Version 8, numeric class levels with no explicit format were ordered by their BEST12. formatted values, and in order to revert to the previous ordering you can specify this format explicitly for the response variable. The change was implemented because the former default behavior for RORDER=FORMATTED often resulted in levels not being ordered numerically and usually required the user to intervene with an explicit format or RORDER=INTERNAL to get the more natural ordering. The following table displays the valid *keywords* and describes how PROC GENMOD interprets them.

| **RORDER=** *keyword* | **Levels Sorted by** |
|---|---|
| DATA | order of appearance in the input data set |
| FORMATTED | external formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value |
| FREQ | descending frequency count; levels with the most observations come first in the order |
| INTERNAL | unformatted value |

By default, RORDER=FORMATTED. For RORDER=FORMATTED and RORDER=INTERNAL, the sort order is machine dependent. The DESCENDING option in the PROC GENMOD statement causes the response variable to be

sorted in the reverse of the order displayed in the previous table. For more information on sorting order, refer to the chapter on the SORT procedure in the *SAS Procedures Guide*.

The NOPRINT option, which suppresses displayed output in other SAS procedures, is not available in the PROC GENMOD statement. However, you can use the Output Delivery System (ODS) to suppress all displayed output, store all output on disk for further analysis, or create SAS data sets from selected output. You can suppress all displayed output with the statement ODS SELECT NONE;, and you can turn displayed output back on with the statement ODS SELECT ALL;. See Table 29.3 on page 1438 for the names of output tables available from PROC GENMOD. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

## BY Statement

**BY** *variables* **;**

You can specify a BY statement with PROC GENMOD to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

Since sorting the data changes the order in which PROC GENMOD reads the data, this can affect the sorting order for the levels of classification variables if you have specified ORDER=DATA in the PROC GENMOD statement. This, in turn, affects specifications in the CONTRAST statement.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the GENMOD procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide.*

## CLASS Statement

**CLASS** *variables* **;**

The CLASS statement specifies the classification variables to be used in the analysis. If the CLASS statement is used, it must appear before the MODEL statement.

Classification variables can be either character or numeric. The procedure uses only the first 16 characters of a character variable value.

By default, class levels are determined from the formatted values of the CLASS variables. Refer to the chapter titled "The FORMAT Procedure" in the *SAS Procedures Guide* and the discussion of the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*. Different sort orders for CLASS variables can be requested by the ORDER= option in the PROC GENMOD statement.

## CONTRAST Statement

**CONTRAST** *'label' effect values* $<$ *,... effect values* $>$ $<$ */options* $>$ **;**

The CONTRAST statement provides a means for obtaining a test for a specified hypothesis concerning the model parameters. This is accomplished by specifying a matrix $\mathbf{L}$ for testing the hypothesis $\mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$. You must be familiar with the details of the model parameterization that PROC GENMOD uses. For more information, see the section "Parameterization Used in PROC GENMOD" on page 1412. Computed statistics are based on the asymptotic chi-square distribution of the likelihood ratio statistic, or the generalized score statistic for GEE models, with degrees of freedom determined by the number of linearly independent rows in the $\mathbf{L}$ matrix. You can request Wald chi-square statistics with the Wald option in the CONTRAST statement.

There is no limit to the number of CONTRAST statements that you can specify, but they must appear after the MODEL statement. Statistics for multiple CONTRAST statements are displayed in a single table.

The following parameters are specified in the CONTRAST statement:

*label* identifies the contrast on the output. A label is required for every contrast specified. Labels can be up to 20 characters and must be enclosed in single quotes.

*effect* identifies an effect that appears in the MODEL statement. The value INTERCEPT or intercept can be used as an effect when an intercept is included in the model. You do not need to include all effects that are included in the MODEL statement.

*values* are constants that are elements of the $\mathbf{L}$ vector associated with the effect.

The rows of $\mathbf{L}'$ are specified in order and are separated by commas. Each row is checked for estimability. If PROC GENMOD finds a contrast to be nonestimable, it displays missing values in corresponding rows in the results. Refer to Searle (1971) for a discussion of estimable functions.

If an effect is not specified in the CONTRAST statement, all of its coefficients in the $\mathbf{L}$ matrix are set to 0. If too many values are specified for an effect, the extra ones are ignored. If too few values are specified, the remaining ones are set to 0.

PROC GENMOD handles missing level combinations of classification variables in the same manner as the GLM and MIXED procedures. Parameters corresponding to missing level combinations are not included in the model. This convention can affect the way in which you specify the $\mathbf{L}$ matrix in your CONTRAST statement.

If the elements of $\mathbf{L}$ are not specified for an effect that contains a specified effect, then the elements of the specified effect are distributed over the levels of the higher-order effect just as the GLM procedure does for its CONTRAST and ESTIMATE statements. For example, suppose that the model contains effects A and B and their interaction A*B. If you specify a CONTRAST statement involving A alone, the $\mathbf{L}$ matrix contains nonzero terms for both A and A*B, since A*B contains A.

If you specify the WALD option, the test of hypothesis is based on a Wald chi-square statistic. If you omit the WALD option, the test statistic computed depends on whether an ordinary generalized linear model or a GEE-type model is specified.

For an ordinary generalized linear model, the CONTRAST statement computes the likelihood ratio statistic. This is defined to be twice the difference between the log likelihood of the model unconstrained by the contrast and the log likelihood with the model fitted under the constraint that the linear function of the parameters defined by the contrast is equal to 0. A *p*-value is computed based on the asymptotic chi-square distribution of the chi-square statistic.

If you specify a GEE model with the REPEATED statement, the test is based on a score statistic. The GEE model is fit under the constraint that the linear function of the parameters defined by the contrast is equal to 0. The score chi-square statistic is computed based on the generalized score function. See the "Generalized Score Statistics" section on page 1428 for more information.

The degrees of freedom is the number of linearly independent constraints implied by the CONTRAST statement, that is, the rank of $\mathbf{L}$.

You can specify the following options after a slash (/).

**E**

requests that the $\mathbf{L}$ matrix be displayed.

**SINGULAR =** *number*

tunes the estimability checking. If $\mathbf{v}$ is a vector, define ABS($\mathbf{v}$) to be the absolute value of the element of $\mathbf{v}$ with the largest absolute value. Define C to be equal to ABS($\mathbf{K}'$) if ABS($\mathbf{K}'$) is greater than 0; otherwise, C equals 1 for a row $\mathbf{K}'$ in the contrast. If ABS($\mathbf{K}' - \mathbf{K}'\mathbf{T}$) is greater than C$*$*number*, then $\mathbf{K}$ is declared nonestimable. $\mathbf{T}$ is the Hermite form matrix $(\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X})$, and $(\mathbf{X}'\mathbf{X})^{-}$ represents a

generalized inverse of the matrix $\mathbf{X}'\mathbf{X}$. The value for $number$ must be between 0 and 1; the default value is 1E−4.

**WALD**
requests that a Wald chi-square statistic be computed for the contrast rather than the default likelihood ratio or score statistic. The Wald statistic for testing $\mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$ is defined by

$$S = (\mathbf{L}'\hat{\boldsymbol{\beta}})'(\mathbf{L}'\boldsymbol{\Sigma}\mathbf{L})^{-}(\mathbf{L}'\hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate and $\boldsymbol{\Sigma}$ is its estimated covariance matrix. The asymptotic distribution of $S$ is $\chi^2_r$, where $r$ is the rank of $\mathbf{L}$. Computed $p$-values are based on this distribution.

If you specify a GEE model with the REPEATED statement, $\boldsymbol{\Sigma}$ is the empirical covariance matrix estimate.

## DEVIANCE Statement

**DEVIANCE** *variable = expression* **;**

You can specify a probability distribution other than those available in PROC GENMOD by using the DEVIANCE and VARIANCE statements. You do not need to specify the DEVIANCE or VARIANCE statements if you use the DIST= MODEL statement option to specify a probability distribution. The *variable* identifies the deviance contribution from a single observation to the procedure, and it must be a valid SAS variable name that does not appear in the input data set. The *expression* can be any arithmetic expression supported by the DATA step language, and it is used to define the functional dependence of the deviance on the mean and the response. You use the automatic variables _MEAN_ and _RESP_ to represent the mean and response in the *expression*.

Alternatively, the deviance function can be defined using programming statements (see the section "Programming Statements" on page 1396) and assigned to a variable, which is then listed as the *expression*. This form is convenient for using complex statements such as if-then-else clauses.

The DEVIANCE statement is ignored unless the VARIANCE statement is also specified.

## ESTIMATE Statement

**ESTIMATE** *'label' effect values . . .* < *options* > **;**

The ESTIMATE statement is similar to a CONTRAST statement, except only one-row $\mathbf{L}'$ matrices are permitted. Each row is checked for estimability. If PROC GENMOD finds a contrast to be nonestimable, it displays missing values in corresponding rows in the results. Refer to Searle (1971) for a discussion of estimable functions.

The actual estimate, $\mathbf{L}'\boldsymbol{\beta}$, its approximate standard error, and its confidence limits are displayed. A Wald chi-square test that $\mathbf{L}'\boldsymbol{\beta} = 0$ is also displayed.

The approximate standard error of the estimate is computed as the square root of $\mathbf{L}'\hat{\boldsymbol{\Sigma}}\mathbf{L}$, where $\hat{\Sigma}$ is the estimated covariance matrix of the parameter estimates. If you specify a GEE model in the REPEATED statement, $\hat{\boldsymbol{\Sigma}}$ is the empirical covariance matrix estimate.

If you specify the EXP option, then $\exp(\mathbf{L}'\boldsymbol{\beta})$, its standard error, and its confidence limits are also displayed.

The construction of the $\mathbf{L}$ vector for an ESTIMATE statement follows the same rules as listed under the CONTRAST statement.

You can specify the following options in the ESTIMATE statement after a slash (/).

**ALPHA=***number*
requests that a confidence interval be constructed with confidence level $1 - number$. The value of *number* must be between 0 and 1; the default value is 0.05.

**E**
requests that the $\mathbf{L}$ matrix coefficients be displayed.

**EXP**
requests that $\exp(\mathbf{L}'\boldsymbol{\beta})$, its standard error, and its confidence limits be computed.

## FREQ Statement

**FREQ** | **FREQUENCY** *variable* **;**

The *variable* in the FREQ statement identifies a variable in the input data set containing the frequency of occurrence of each observation. PROC GENMOD treats each observation as if it appears $n$ times, where $n$ is the value of the FREQ variable for the observation. If it is not an integer, the frequency value is truncated to an integer. If it is less than 1 or if it is missing, the observation is not used.

## FWDLINK Statement

**FWDLINK** *variable = expression* **;**

You can define a link function other than a built-in link function by using the FWDLINK statement. If you use the MODEL statement option LINK= to specify a link function, you do not need to use the FWDLINK statement. The *variable* identifies the link function to the procedure. The *expression* can be any arithmetic expression supported by the DATA step language, and it is used to define the functional dependence on the mean.

Alternatively, the link function can be defined by using programming statements (see the "Programming Statements" section on page 1396) and assigned to a variable,

which is then listed as the *expression*. The second form is convenient for using complex statements such as if-then-else clauses. The GENMOD procedure automatically computes derivatives of the link function required for iterative fitting. You must specify the inverse of the link function in the INVLINK statement when you specify the FWDLINK statement to define the link function. You use the automatic variable _MEAN_ to represent the mean in the preceding *expression*.

## INVLINK Statement

**INVLINK** *variable = expression* **;**

If you define a link function in the FWDLINK statement, then you must define the inverse link function using the INVLINK statement. If you use the MODEL statement option LINK= to specify a link function, you do not need to use the INVLINK statement. The *variable* identifies the inverse link function to the procedure. The *expression* can be any arithmetic expression supported by the DATA step language, and it is used to define the functional dependence on the linear predictor.

Alternatively, the inverse link function can be defined using programming statements (see the section "Programming Statements" on page 1396) and assigned to a variable, which is then listed as the *expression*. The second form is convenient for using complex statements such as if-then-else clauses. The automatic variable _XBETA_ represents the linear predictor in the preceding *expression*.

## LSMEANS Statement

**LSMEANS** *effects* < */ options* > **;**

The LSMEANS statement computes least-squares means (LS-means) corresponding to the specified effects for the linear predictor part of the model. The $\mathbf{L}$ matrix constructed to compute them is precisely the same as the one formed in PROC GLM.

The LSMEANS statement is not available for multinomial distribution models for ordinal response data.

Each LS-mean is computed as $\mathbf{L}'\widehat{\boldsymbol{\beta}}$, where $\mathbf{L}$ is the coefficient matrix associated with the least-squares mean and $\widehat{\boldsymbol{\beta}}$ is the estimate of the parameter vector. The approximate standard errors for the LS-mean is computed as the square root of $\mathbf{L}'\boldsymbol{\Sigma}\mathbf{L}$.

LS-means can be computed for any effect in the MODEL statement that involves CLASS variables. You can specify multiple effects in one LSMEANS statement or multiple LSMEANS statements, and all LSMEANS statements must appear after the MODEL statement.

As in the ESTIMATE statement, the **L** matrix is tested for estimability, and if this test fails, PROC GENMOD displays "Non-est" for the LS-means entries.

Assuming the LS-mean is estimable, PROC GENMOD constructs a Wald chi-square test to test the null hypothesis that the associated population quantity equals zero.

You can specify the following options in the LSMEANS statement after a slash (/).

**ALPHA=***number*
requests that a confidence interval be constructed for each of the LS-means with confidence level $(1 - number) \times 100\%$. The value of *number* must be between 0 and 1; the default value is 0.05, corresponding to a 95% confidence interval.

**CL**
requests that confidence limits be constructed for each of the LS-means. The confidence level is 0.95 by default; this can be changed with the ALPHA= option.

**CORR**
displays the estimated correlation matrix of the LS-means as part of the "Least Squares Means" table.

**COV**
displays the estimated covariance matrix of the LS-means as part of the "Least Squares Means" table.

**DIFF**
requests that differences of the LS-means be displayed. All possible differences of LS-means, standard errors, and a Wald chi-square test are computed. Confidence limits are computed if the CL option is also specified.

**E**
requests that the **L** matrix coefficients for all LSMEANS effects be displayed.

## MAKE Statement

**MAKE***'table'* **OUT=***SAS-data-set***;**

PROC GENMOD assigns a name to each table that it creates. The MAKE statement creates a SAS data set containing the results in the table named as 'table'. The MAKE statement is included for compatibility with GENMOD in SAS release 6.12.

You can use a table name to reference a table when using either the MAKE statement or the Output Delivery System (ODS) to create output data sets. ODS is the recommended method to create SAS data sets from displayed output, since other SAS procedures use it in SAS release 7 and later, and it is more flexible than the MAKE statement. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

## MODEL Statement

**MODEL** *response* = < *effects* >< */options* > **;**
**MODEL** *events/trials* = < *effects* >< */options* > **;**

The MODEL statement specifies the response, or dependent variable, and the effects, or explanatory variables. If you omit the explanatory variables, the procedure fits an intercept-only model. An intercept term is included in the model by default. The intercept can be removed with the NOINT option.

You can specify the response in the form of a single variable or in the form of a ratio of two variables denoted *events/trials*. The first form is applicable to all responses. The second form is applicable only to summarized binomial response data. When each observation in the input data set contains the number of events (for example, successes) and the number of trials from a set of binomial trials, use the *events/trials* syntax.

In the *events/trials* model syntax, you specify two variables that contain the event and trial counts. These two variables are separated by a slash (/). The values of both *events* and (*trials*−*events*) must be nonnegative, and the value of the *trials* variable must be greater than 0 for an observation to be valid. The variable *events* or *trials* may take noninteger values.

When each observation in the input data set contains a single trial from a binomial or multinomial experiment, use the first form of the MODEL statement above. The response variable can be numeric or character. The ordering of response levels is critical in these models. You can use the RORDER= option in the PROC GENMOD statement to specify the response level ordering.

Responses for the Poisson distribution must be positive, but they can be noninteger values.

The effects in the MODEL statement consist of an explanatory variable or combination of variables. Explanatory variables can be continuous or classification variables. Classification variables can be character or numeric. Explanatory variables representing nominal, or classification, data must be declared in a CLASS statement. Interactions between variables can also be included as effects. Columns of the design matrix are automatically generated for classification variables and interactions. The syntax for specification of effects is the same as for the GLM procedure. See the “Specification of Effects” section on page 1411 for more information. Also refer to Chapter 30, “The GLM Procedure.”

You can specify the following options in the MODEL statement after a slash (/).

**AGGREGATE=** *(variable-list)*
**AGGREGATE=** *variable*
specifies the subpopulations on which the Pearson chi-square and the deviance are calculated. This option applies only to the multinomial distribution or the binomial distribution with binary (single trial syntax) response. It is ignored if specified for other cases. Observations with common values in the given list of variables are regarded as coming from the same subpopulation. This affects the computation of the deviance and Pearson chi-square statistics. Variables in the list can be any variables in the input data set.

**ALPHA** | **ALPH** | **A=***number*
sets the confidence coefficient for parameter confidence intervals to $1-$*number*. The value of *number* must be between 0 and 1. The default value of *number* is 0.05.

**CICONV=***number*
sets the convergence criterion for profile likelihood confidence intervals. See the section "Confidence Intervals for Parameters" on page 1415 for the definition of convergence. The value of *number* must be between 0 and 1. By default, CICONV=1E$-$4.

**CL**
requests that confidence limits for predicted values be displayed. See the OBSTATS option.

**CONVERGE=***number*
sets the convergence criterion. The value of *number* must be between 0 and 1. The iterations are considered to have converged when the maximum change in the parameter estimates between iteration steps is less than the value specified. The change is a relative change if the parameter is greater than 0.01 in absolute value; otherwise, it is an absolute change. By default, CONVERGE=1E$-$4. This convergence criterion is used in parameter estimation for a single model fit, Type 1 statistics, and likelihood ratio statistics for Type 3 analyses and CONTRAST statements.

**CONVH=***number*
sets the relative Hessian convergence criterion. The value of *number* must be between 0 and 1. After convergence is determined with the change in parameter criterion specified with the CONVERGE= option, the quantity $tc = \frac{\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}}{|f|}$ is computed and compared to *number*, where $\mathbf{g}$ is the gradient vector, $\mathbf{H}$ is the Hessian matrix for the model parameters, and $f$ is the log-likelihood function. If $tc$ is greater than *number*, a warning that the relative Hessian convergence criterion has been exceeded is printed. This criterion detects the occasional case where the change in parameter convergence criterion is satisfied, but a maximum in the log-likelihood function has not been attained. By default, CONVH=1E$-$4.

**CORRB**
requests that the parameter estimate correlation matrix be displayed.

**COVB**
requests that the parameter estimate covariance matrix be displayed.

**DIST | D | ERROR | ERR =** *keyword*

specifies the built-in probability distribution to use in the model. If you specify the DIST= option and you omit a user-defined link function, a default link function is chosen as displayed in the following table. If you specify no distribution and no link function, then the GENMOD procedure defaults to the normal distribution with the identity link function.

| DIST= | Distribution | Default Link Function |
|---|---|---|
| BINOMIAL \| BIN \| B | binomial | logit |
| GAMMA \| GAM \| G | gamma | inverse ( power(−1) ) |
| IGAUSSIAN \| IG | inverse Gaussian | inverse squared ( power(−2) ) |
| MULTINOMIAL \| MULT | multinomial | cumulative logit |
| NEGBIN \| NB | negative binomial | log |
| NORMAL \| NOR \| N | normal | identity |
| POISSON \| POI \| P | Poisson | log |

**EXPECTED**

requests that the expected Fisher information matrix be used to compute parameter estimate covariances and the associated statistics. The default action is to use the observed Fisher information matrix. See the SCORING= option.

**ID=***variable*

causes the values of *variable* in the input data set to be displayed in the OBSTATS table. If an explicit format for *variable* has been defined, the formatted values are displayed. If the OBSTATS option is not specified, this option has no effect.

**INITIAL=***numbers*

sets initial values for parameter estimates in the model. The default initial parameter values are weighted least squares estimates based on using the response data as the initial mean estimate. This option can be useful in case of convergence difficulty. The intercept parameter is initialized with the INTERCEPT= option and is not included here. The values are assigned to the variables in the MODEL statement in the same order in which they appear in the MODEL statement. The order of levels for CLASS variables is determined by the ORDER= option. Note that some levels of class variables can be aliased; that is, they correspond to linearly dependent parameters that are not estimated by the procedure. Initial values must be assigned to all levels of class variables, regardless of whether they are aliased or not. The procedure ignores initial values corresponding to parameters not being estimated. If you specify a BY statement, all class variables must take on the same number of levels in each BY group. Otherwise, class variables in some of the BY groups are assigned incorrect initial values. Types of INITIAL= specifications are illustrated in the following table.

| Type of List | Specification |
|---|---|
| list separated by blanks | INITIAL = 3 4 5 |
| list separated by commas | INITIAL = 3, 4, 5 |
| x to y | INITIAL = 3 to 5 |
| x to y by z | INITIAL = 3 to 5 by 1 |
| combination of list types | INITIAL = 1, 3 to 5, 9 |

**INTERCEPT=***number*

initializes the intercept term to *number* for parameter estimation. If you specify both the INTERCEPT= and the NOINT options, the intercept term is not estimated, but an intercept term of *number* is included in the model.

**ITPRINT**

displays the iteration history for all iterative processes: parameter estimation, fitting constrained models for contrasts and Type 3 analyses, and profile likelihood confidence intervals. The last evaluation of the gradient and the negative of the Hessian (second derivative) matrix are also displayed for parameter estimation. This option may result in a large amount of displayed output, especially if some of the optional iterative processes are selected.

**LINK =** *keyword*

specifies the link function to use in the model. The keywords and their associated built-in link functions are as follows.

| **LINK=** | **Link Function** |
|---|---|
| CUMCLL \| CCLL | cumulative complementary log-log |
| CUMLOGIT \| CLOGIT | cumulative logit |
| CUMPROBIT \| CPROBIT | cumulative probit |
| CLOGLOG \| CLL | complementary log-log |
| IDENTITY \| ID | identity |
| LOG | log |
| LOGIT | logit |
| PROBIT | probit |
| POWER(*number*) \| POW(*number*) | power with $\lambda$= *number* |

If no LINK= option is supplied and there is a user-defined link function, the user-defined link function is used. If you specify neither the LINK= option nor a user-defined link function, then the default canonical link function is used if you specify the DIST= option. Otherwise, if you omit the DIST= option, the identity link function is used.

The cumulative link functions are appropriate only for the multinomial distribution.

**LRCI**

requests that two-sided confidence intervals for all model parameters be computed based on the profile likelihood function. This is sometimes called the partially maximized likelihood function. See the "Confidence Intervals for Parameters" section on page 1415 for more information on the profile likelihood function. This computation is iterative and can consume a relatively large amount of CPU time. The confidence coefficient can be selected with the ALPHA=*number* option. The resulting confidence coefficient is 1−*number*. The default confidence coefficient is 0.95.

**MAXITER=***number*
**MAXIT=***number*

sets the maximum allowable number of iterations for all iterative computation processes in PROC GENMOD. By default, MAXITER=50.

**NOINT**
requests that no intercept term be included in the model. An intercept is included unless this option is specified.

**NOSCALE**
holds the scale parameter fixed. Otherwise, for the normal, inverse gaussian, and gamma distributions, the scale parameter is estimated by maximum likelihood. If you omit the SCALE= option, the scale parameter is fixed at the value 1.

**OFFSET=***variable*
specifies a variable in the input data set to be used as an offset variable. This variable cannot be a CLASS variable, and it cannot be the response variable or one of the explanatory variables.

**OBSTATS**
specifies that an additional table of statistics be displayed. For each observation, the following items are displayed:

- the value of the response variable (variables if the data are binomial), frequency, and weight variables
- the values of the regression variables
- predicted mean, $\hat{\mu} = g^{-1}(\eta)$, where $\eta = \mathbf{x}_i{}'\hat{\boldsymbol{\beta}}$ is the linear predictor and $g$ is the link function. If there is an offset, it is included in $\mathbf{x}_i{}'\hat{\boldsymbol{\beta}}$.
- estimate of the linear predictor $\mathbf{x}_i{}'\hat{\boldsymbol{\beta}}$. If there is an offset, it is included in $\mathbf{x}_i{}'\hat{\boldsymbol{\beta}}$.
- standard error of the linear predictor $\mathbf{x}_i{}'\hat{\boldsymbol{\beta}}$
- the value of the Hessian weight at the final iteration
- lower confidence limit of the predicted value of the mean. The confidence coefficient is specified with the ALPHA= option. See the section "Confidence Intervals on Predicted Values" on page 1417 for the computational method.
- upper confidence limit of the predicted value of the mean
- raw residual, defined as $Y - \mu$
- Pearson, or chi residual, defined as the square root of the contribution for the observation to the Pearson chi-square, that is

$$\frac{Y - \mu}{\sqrt{V(\mu)/w}}$$

  where $Y$ is the response, $\mu$ is the predicted mean, $w$ is the value of the prior weight variable specified in a WEIGHT statement, and V($\mu$) is the variance function evaluated at $\mu$.
- the standardized Pearson residual
- deviance residual, defined as the square root of the deviance contribution for the observation, with sign equal to the sign of the raw residual
- the standardized deviance residual
- the likelihood residual

The RESIDUALS, PREDICTED, XVARS, and CL options cause only subgroups of the observation statistics to be displayed. You can specify more than one of these options to include different subgroups of statistics.

The ID=*variable* option causes the values of *variable* in the input data set to be displayed in the table. If an explicit format for *variable* has been defined, the formatted values are displayed.

If a REPEATED statement is present, a table is displayed for the GEE model specified in the REPEATED statement. Only the regression variables, response values, predicted values, confidence limits for the predicted values, linear predictor, raw residuals, and Pearson residuals for each observation in the input data set are available.

**PREDICTED**
**PRED**
**P**

requests that predicted values, the linear predictor, its standard error, and the Hessian weight be displayed. See the OBSTATS option.

**RESIDUALS**
**R**

requests that residuals and standardized residuals be displayed. See the OBSTATS option.

**SCALE=***number*
**SCALE=PEARSON**
**SCALE=P**
**PSCALE**
**SCALE=DEVIANCE**
**SCALE=D**
**DSCALE**

sets the value used for the scale parameter where the NOSCALE option is used. For the binomial and Poisson distributions, which have no free scale parameter, this can be used to specify an *overdispersed* model. In this case, the parameter covariance matrix and the likelihood function are adjusted by the scale parameter. See the "Dispersion Parameter" section (page 1409) and the "Overdispersion" section (page 1410) for more information. If the NOSCALE option is not specified, then *number* is used as an initial estimate of the scale parameter.

Specifying SCALE=PEARSON or SCALE=P is the same as specifying the PSCALE option. This fixes the scale parameter at the value 1 in the estimation procedure. After the parameter estimates are determined, the exponential family dispersion parameter is assumed to be given by Pearson's chi-square statistic divided by the degrees of freedom, and all statistics such as standard errors and likelihood ratio statistics are adjusted appropriately.

Specifying SCALE=DEVIANCE or SCALE=D is the same as specifying the DSCALE option. This fixes the scale parameter at a value of 1 in the estimation procedure.

After the parameter estimates are determined, the exponential family dispersion parameter is assumed to be given by the deviance divided by the degrees of freedom. All statistics such as standard errors and likelihood ratio statistics are adjusted appropriately.

**SCORING=***number*

requests that on iterations up to *number*, the Hessian matrix is computed using the Fisher's scoring method. For further iterations, the full Hessian matrix is computed. The default value is 1. A value of 0 causes all iterations to use the full Hessian matrix, and a value greater than or equal to the value of the MAXITER option causes all iterations to use Fisher's scoring. The value of the SCORING= option must be 0 or a positive integer.

**SINGULAR=***number*

sets the tolerance for testing singularity of the information matrix and the crossproducts matrix. Roughly, the test requires that a pivot be at least this number times the original diagonal value. By default, *number* is $10^7$ times the machine epsilon. The default *number* is approximately $10^{-9}$ on most machines.

**TYPE1**

requests that a Type 1, or sequential, analysis be performed. This consists of sequentially fitting models, beginning with the null (intercept term only) model and continuing up to the model specified in the MODEL statement. The likelihood ratio statistic between each successive pair of models is computed and displayed in a table.

A Type 1 analysis is not available for GEE models, since there is no associated likelihood.

**TYPE3**

requests that statistics for Type 3 contrasts be computed for each effect specified in the MODEL statement. The default analysis is to compute likelihood ratio statistics for the contrasts or score statistics for GEEs. Wald statistics are computed if the WALD option is also specified.

**WALD**

requests Wald statistics for Type 3 contrasts. You must also specify the TYPE3 option in order to compute Type 3 Wald statistics.

**WALDCI**

requests that two-sided Wald confidence intervals for all model parameters be computed based on the asymptotic normality of the parameter estimators. This computation is not as time consuming as the LRCI method, since it does not involve an iterative procedure. However, it is not thought to be as accurate, especially for small sample sizes. The confidence coefficient can be selected with the ALPHA= option in the same way as for the LRCI option.

**XVARS**

requests that the regression variables be included in the OBSTATS table.

## OUTPUT Statement

**OUTPUT**< **OUT=***SAS-data-set* >
< *keyword=name . . . keyword=name* > / **;**

The OUTPUT statement creates a new SAS data set that contains all the variables in the input data set and, optionally, the estimated linear predictors (XBETA) and their standard error estimates, the weights for the Hessian matrix, predicted values of the mean, confidence limits for predicted values, and residuals.

You can also request these statistics with the OBSTATS, PREDICTED, RESIDUALS, CL, or XVARS options in the MODEL statement. You can then create a SAS data set containing them with ODS OUTPUT commands. You may prefer to specify the OUTPUT statement for requesting these statistics since

- the OUTPUT statement produces no tabular output
- the OUTPUT statement creates a SAS data set more efficiently than ODS. This can be an advantage for large data sets.
- you can specify the individual statistics to be included in the SAS data set

If you use the multinomial distribution with one of the cumulative link functions for ordinal data, the data set also contains variables named _ORDER_ and _LEVEL_ that indicate the levels of the ordinal response variable and the values of the variable in the input data set corresponding to the sorted levels. These variables indicate that the predicted value for a given observation is the probability that the response variable is as large as the value of the Value variable.

The estimated linear predictor, its standard error estimate, and the predicted values and their confidence intervals are computed for all observations in which the explanatory variables are all nonmissing, even if the response is missing. By adding observations with missing response values to the input data set, you can compute these statistics for new observations or for settings of the explanatory variables not present in the data without affecting the model fit.

The following list explains specifications in the OUTPUT statement.

**OUT=** *SAS-data-set*
specifies the output data set. If you omit the OUT=option, the output data set is created and given a default name using the DATA$n$ convention.

*keyword=name*
specifies the statistics to be included in the output data set and names the new variables that contain the statistics. Specify a keyword for each desired statistic (see the following list of keywords), an equal sign, and the name of the new variable or variables to contain the statistic. You can list only one variable after the equal sign. Although you can use the OUTPUT statement without any *keyword=name* specifications, the output data set then contains only the original variables and, possibly, the variables Level and Value (if you use the multinomial model with ordinal data).

Note that the residuals are not available for the multinomial model with ordinal data. Formulas for the statistics are given in the section "Predicted Values of the Mean" on page 1417 and the "Residuals" section on page 1418. The keywords allowed and the statistics they represent are as follows:

HESSWGT — diagonal element of the weight matrix used in computing the Hessian matrix

LOWER | L — lower confidence limit for the predicted value of the mean, or the lower confidence limit for the probability that the response is less than or equal to the value of Level or Value. The confidence coefficient is determined by the ALPHA=$number$ option in the MODEL statement as $(1 - number) \times 100\%$. The default confidence coefficient is 95%.

PREDICTED | PRED | PROB | P — predicted value of the mean or the predicted probability that the response variable is less than or equal to the value of Level or Value if the multinomial model for ordinal data is used (in other words, $\Pr(\mathrm{Y} \leq Value)$, where Y is the response variable)

RESCHI — Pearson (Chi) residual for identifying observations that are poorly accounted for by the model

RESDEV — deviance residual for identifying poorly fitted observations

RESLIK — likelihood residual for identifying poorly fitted observations

STDXBETA — standard error estimate of XBETA (see the XBETA keyword)

STDRESCHI — standardized Pearson (Chi) residual for identifying observations that are poorly accounted for by the model

STDRESDEV — standardized deviance residual for identifying poorly fitted observations

UPPER | U — upper confidence limit for the predicted value of the mean, or the lower confidence limit for the probability that the response is less than or equal to the value of Level or Value. The confidence coefficient is determined by the ALPHA=$number$ option in the MODEL statement as $(1 - number) \times 100\%$. The default confidence coefficient is 95%.

XBETA — estimate of the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ for observation $i$, or $\alpha_j + \mathbf{x}_i'\boldsymbol{\beta}$, where $j$ is the corresponding ordered value of the response variable for the multinomial model with ordinal data. If there is an offset, it is included in $\mathbf{x}_i'\boldsymbol{\beta}$.

## Programming Statements

Although the most commonly used link and probability distributions are available as built-in functions, the GENMOD procedure enables you to define your own link functions and response probability distributions using the FWDLINK, INVLINK, VARIANCE, and DEVIANCE statements. The variables assigned in these statements can have values computed in programming statements. These programming statements

can occur anywhere between the PROC GENMOD statement and the RUN statement. Variable names used in programming statements must be unique. Variables from the input data set may be referenced in programming statements. The mean, linear predictor, and response are represented by the automatic variables _MEAN_, _XBETA_, and _RESP_, which can be referenced in your programming statements. Programming statements are used to define the functional dependencies of the link function, the inverse link function, the variance function, and the deviance function on the mean, linear predictor, and response variable.

The following code illustrates the use of programming statements. Even though you usually request the Poisson distribution by specifying DIST=POISSON as a MODEL statement option, you can define the variance and deviance functions for the Poisson distribution by using the VARIANCE and DEVIANCE statements. For example, the following code performs the same analysis as the Poisson regression example in the "Getting Started" section on page 1370. The code must be in logical order for computation, just as in a DATA step.

```
proc genmod ;
   class car age;
   a = _MEAN_;
   y = _RESP_;
   d = 2 * ( y * log( y / a ) - ( y - a ) );
   variance var = a;
   deviance dev = d;
   model c = car age / link = log offset = ln;
run;
```

The variables var and dev are dummy variables used internally by the procedure to identify the variance and deviance functions. Any valid SAS variable names can be used.

Similarly, the log link function and its inverse could be defined with the FWDLINK and INVLINK statements.

```
fwdlink link = log(_MEAN_);
invlink ilink = exp(_XBETA_);
```

This code is for illustration, and it works well for most Poisson regression problems. If, however, in the iterative fitting process, the mean parameter becomes too close to 0, or a 0 response value occurs, an error condition occurs when the procedure attempts to evaluate the log function. You can circumvent this kind of problem by using if-then-else clauses or other conditional statements to check for possible error conditions and appropriately define the functions for these cases.

Data set variables can be referenced in user definitions of the link function and response distributions using programming statements and the FWDLINK, INVLINK, DEVIANCE, and VARIANCE statements.

See the DEVIANCE, VARIANCE, FWDLINK, and INVLINK statements for more information.

## REPEATED Statement

**REPEATED SUBJECT=** *subject-effect* < */ options* > **;**

The REPEATED statement specifies the covariance structure of multivariate responses for GEE model fitting in the GENMOD procedure. In addition, the REPEATED statement controls the iterative fitting algorithm used in GEEs and specifies optional output. Other GENMOD procedure statements, such as the MODEL and CLASS statements, are used in the same way as they are for ordinary generalized linear models to specify the regression model for the mean of the responses.

**SUBJECT=***subject-effect*
identifies subjects in the input data set. The *subject-effect* can be a single variable, an interaction effect, a nested effect, or a combination. Each distinct value, or level, of the effect identifies a different subject, or cluster. Responses from different subjects are assumed to be statistically independent, and responses within subjects are assumed to be correlated. A *subject-effect* must be specified, and variables used in defining the *subject-effect* must be listed in the CLASS statement. The input data set does not need to be sorted by subject. See the SORTED option.

The *options* control how the model is fit and what output is produced. You can specify the following options after a slash (/).

**ALPHAINIT=***numbers*
specifies initial values for log odds ratio regression parameters if the LOGOR= option is specified for binary data. If this option is not specified, an initial value of 0.01 is used for all the parameters.

**CONVERGE=***number*
specifies the convergence criterion for GEE parameter estimation. If the maximum absolute difference between regression parameter estimates is less than the value of *number* on two successive iterations, convergence is declared. If the absolute value of a regression parameter estimate is greater than 0.08, then the absolute difference normalized by the regression parameter value is used instead of the absolute difference. The default value of *number* is 0.0001.

**CORRW**
displays the estimated working correlation matrix.

**CORRB**
displays the estimated regression parameter correlation matrix. Both model-based and empirical correlations are displayed.

**COVB**
displays the estimated regression parameter covariance matrix. Both model-based and empirical covariances are displayed.

**ECORRB**
displays the estimated regression parameter empirical correlation matrix.

**ECOVB**
displays the estimated regression parameter empirical covariance matrix.

**INTERCEPT=***number*
specifies either an initial or a fixed value of the intercept regression parameter in the GEE model. If you specify the NOINT option in the MODEL statement, then the intercept is fixed at the value of *number*.

**INITIAL=***numbers*
specifies initial values of the regression parameters estimation, other than the intercept parameter, for GEE estimation. If this option is not specified, the estimated regression parameters assuming independence for all responses are used for the initial values.

**LOGOR=***log odds ratio structure keyword*
specifies the regression structure of the log odds ratio used to model the association of the responses from subjects for binary data. The response syntax must be of the single variable type, the distribution must be binomial, and the data must be binary. The following table displays the log odds ratio structure keywords and the corresponding log odds ratio regression structures. See the "Alternating Logistic Regressions" section on page 1424 for definitions of the log odds ratio types and examples of specifying log odds ratio models. You should specify either the LOGOR= or the TYPE= option, but not both.

**Table 29.1.** Log Odds Ratio Regression Structures

| **Keyword** | **Log Odds Ratio Regression Structure** |
|---|---|
| EXCH | exchangeable |
| FULLCLLUST | fully parameterized clusters |
| LOGORVAR(*variable*) | indicator variable for specifying block effects |
| NESTK | *k*-nested |
| NEST1 | 1-nested |
| ZFULL | fully specified *z*-matrix specified in ZDATA= data set |
| ZREP | single cluster specification for replicated *z*-matrix specified in ZDATA= data set |
| ZREP(matrix) | single cluster specification for replicated *z*-matrix |

**MAXITER=***number*
**MAXIT=***number*
specifies the maximum number of iterations allowed in the iterative GEE estimation process. The default number is 50.

**MCORRB**
displays the estimated regression parameter model-based correlation matrix.

**MCOVB**
displays the estimated regression parameter model-based covariance matrix.

**MODELSE**
displays an analysis of parameter estimates table using model-based standard errors. By default, an "Analysis of Parameter Estimates" table based on empirical standard errors is displayed.

**RUPDATE=***number*
specifies the number of iterations between updates of the working correlation matrix. For example, RUPDATE=5 specifies that the working correlation is updated once for every five regression parameter updates. The default value of *number* is 1; that is, the working correlation is updated every time the regression parameters are updated.

**SORTED**
specifies that the input data are grouped by subject and sorted within subject. If this option is not specified, then the procedure internally sorts by *subject-effect* and *within subject-effect*, if a *within subject-effect* is specified.

**SUBCLUSTER=***variable*
**SUBCLUST=***variable*
specifies a variable defining subclusters for the 1-nested or *k*-nested log odds ratio association modeling structures.

**TYPE | CORR=***correlation-structure keyword*
specifies the structure of the working correlation matrix used to model the correlation of the responses from subjects. The following table displays the correlation structure keywords and the corresponding correlation structures. The default working correlation type is the independent (CORR=IND). See the "Details" section on page 1402 for definitions of the correlation matrix types. You should specify LOGOR= or TYPE= but not both.

**Table 29.2.** Correlation Structure Types

| Keyword | Correlation Matrix Type |
|---|---|
| AR \| AR(1) | autoregressive(1) |
| EXCH \| CS | exchangeable |
| IND | independent |
| MDEP(number) | $m$-dependent with $m$=number |
| UNSTR \| UN | unstructured |
| USER \| FIXED (matrix) | fixed, user-specified correlation matrix |

For example, you can specify a fixed $4 \times 4$ correlation matrix with the option

```
TYPE=USER( 1.0  0.9  0.8  0.6
           0.9  1.0  0.9  0.8
           0.8  0.9  1.0  0.9
           0.6  0.8  0.9  1.0 )
```

**V6CORR**
specifies that the 'Version 6' method of computing the normalized Pearson chi-square be used for working correlation estimation and for model-based covariance matrix scale factor.

**WITHINSUBJECT | WITHIN=***within subject-effect*
defines an effect specifying the order of measurements within subjects. Each distinct level of the *within subject-effect* defines a different response from the same subject. If the data are in proper order within each subject, you do not need to specify this option.

If some measurements do not appear in the data for some subjects, this option properly orders the existing measurements and treats the omitted measurements as missing values. If the WITHINSUBJECT= option is not used in this situation, measurements may be improperly ordered and missing values assumed for the last measurements in a cluster.

Variables used in defining the *within subject-effect* must be listed in the CLASS statement.

**YPAIR=***variable-list*
specifies the variables in the ZDATA= data set corresponding to pairs of responses for log odds ratio association modeling.

**ZDATA=***SAS-data-set*
specifies a SAS data set containing either the full *z*-matrix for log odds ratio association modeling or the *z*-matrix for a single complete cluster to be replicated for all clusters.

**ZROW=***variable-list*
specifies the variables in the ZDATA= data set corresponding to rows of the *z*-matrix for log odds ratio association modeling.

## VARIANCE Statement

**VARIANCE** *variable = expression* **;**

You can specify a probability distribution other than the built-in distributions by using the VARIANCE and DEVIANCE statements. The variable name *variable* identifies the variance function to the procedure. The *expression* is used to define the functional dependence on the mean, and it can be any arithmetic expression supported by the DATA step language. You use the automatic variable _MEAN_ to represent the mean in the expression.

Alternatively, you can define the variance function with programming statements, as detailed in the section "Programming Statements" on page 1396. This form is convenient for using complex statements such as if-then-else clauses. Derivatives of the variance function for use during optimization are computed automatically. The DEVIANCE statement must also appear when the VARIANCE statement is used to define the variance function.

## WEIGHT Statement

**WEIGHT | SCWGT** *variable* **;**

The WEIGHT statement identifies a variable in the input data set to be used as the exponential family dispersion parameter weight for each observation. The exponential family dispersion parameter is divided by the WEIGHT variable value for each observation. This is true regardless of whether the parameter is estimated by the procedure or specified in the MODEL statement with the SCALE= option. It is also true for distributions such as the Poisson and binomial that are not usually defined to have a dispersion parameter. For these distributions, a WEIGHT variable weights the overdispersion parameter, which has the default value of 1.

The WEIGHT variable does not have to be an integer; if it is less than or equal to 0 or if it is missing, the corresponding observation is not used.

# Details

## Generalized Linear Models Theory

This is a brief introduction to the theory of generalized linear models . See the "References" section on page 1462 for sources of more detailed information.

### *Response Probability Distributions*

In generalized linear models, the response is assumed to possess a probability distribution of the exponential form. That is, the probability density of the response $Y$ for continuous response variables, or the probability function for discrete responses, can be expressed as

$$f(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

for some functions $a$, $b$, and $c$ that determine the specific distribution. For fixed $\phi$, this is a one parameter exponential family of distributions. The functions $a$ and $c$ are such that $a(\phi) = \phi/w$ and $c = c(y, \phi/w)$, where $w$ is a known weight for each observation. A variable representing $w$ in the input data set may be specified in the WEIGHT statement. If no WEIGHT statement is specified, $w_i = 1$ for all observations.

Standard theory for this type of distribution gives expressions for the mean and variance of $Y$.

$$\begin{aligned} E(Y) &= b'(\theta) \\ \mathit{Var}(Y) &= \frac{b''(\theta)\phi}{w} \end{aligned}$$

where the primes denote derivatives with respect to $\theta$. If $\mu$ represents the mean of $Y$, then the variance expressed as a function of the mean is

$$\mathit{Var}(Y) = \frac{V(\mu)\phi}{w}$$

where $V$ is the *variance function*.

Probability distributions of the response $Y$ in generalized linear models are usually parameterized in terms of the mean $\mu$ and dispersion parameter $\phi$ instead of the *natural parameter* $\theta$. The probability distributions that are available in the GENMOD procedure are shown in the following list. The PROC GENMOD scale parameter and the variance of $Y$ are also shown.

- Normal:

$$\begin{aligned} f(y) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] \quad \text{for } -\infty < y < \infty \\ \phi &= \sigma^2 \\ \mathit{scale} &= \sigma \\ Var(Y) &= \sigma^2 \end{aligned}$$

- Inverse Gaussian:

$$\begin{aligned} f(y) &= \frac{1}{\sqrt{2\pi y^3}\sigma}\exp\left[-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right] \quad \text{for } 0 < y < \infty \\ \phi &= \sigma^2 \\ \mathit{scale} &= \sigma \\ \mathit{Var}(Y) &= \sigma^2\mu^3 \end{aligned}$$

- Gamma:

$$\begin{aligned} f(y) &= \frac{1}{\Gamma(\nu)y}\left(\frac{y\nu}{\mu}\right)^{\nu}\exp\left(-\frac{y\nu}{\mu}\right) \quad \text{for } 0 < y < \infty \\ \phi &= \nu^{-1} \\ \textit{scale} &= \nu \\ \textit{Var}(Y) &= \frac{\mu^2}{\nu} \end{aligned}$$

- Negative Binomial:

$$\begin{aligned} f(y) &= \frac{\Gamma(y+1/k)}{\Gamma(y+1)\Gamma(1/k)}\frac{(k\mu)^k}{(1+k\mu)^{y+1/k}} \quad \text{for } y = 0, 1, 2, \ldots \\ \textit{dispersion} &= k \\ \textit{Var}(Y) &= \mu + k\mu^2 \end{aligned}$$

- Poisson:

$$\begin{aligned} f(y) &= \frac{\mu^y e^{-\mu}}{y!} \quad \text{for } y = 0, 1, 2, \ldots \\ \phi &= 1 \\ \textit{Var}(Y) &= \mu \end{aligned}$$

- Binomial:

$$\begin{aligned} f(y) &= \binom{n}{r}\mu^r(1-\mu)^{n-r} \quad \text{for } y = \frac{r}{n},\ r = 0, 1, 2, \ldots, n \\ \phi &= 1 \\ \textit{Var}(Y) &= \frac{\mu(1-\mu)}{n} \end{aligned}$$

- Multinomial:

$$f(y_1, y_2, \cdots, y_k) = \frac{m!}{y_1!y_2!\cdots y_k!}p_1^{y_1}p_2^{y_2}\cdots p_k^{y_k}$$

The negative binomial distribution contains a parameter $k$, called the negative binomial dispersion parameter. This is not the same as the generalized linear model dispersion $\phi$, but it is an additional distribution parameter that must be estimated or set to a fixed value.

For the binomial distribution, the response is the binomial proportion $Y = \textit{events}/\textit{trials}$. The variance function is $V(\mu) = \mu(1-\mu)$, and the binomial trials parameter $n$ is regarded as a weight $w$.

If a weight variable is present, $\phi$ is replaced with $\phi/w$, where $w$ is the weight variable.

PROC GENMOD works with a scale parameter that is related to the exponential family dispersion parameter $\phi$ instead of with $\phi$ itself. The scale parameters are related to the dispersion parameter as shown previously with the probability distribution definitions. Thus, the scale parameter output in the "Analysis of Parameter Estimates" table is related to the exponential family dispersion parameter. If you specify a constant scale parameter with the SCALE= option in the MODEL statement, it is also related to the exponential family dispersion parameter in the same way.

### Link Function

The mean $\mu_i$ of the response in the $i$th observation is related to a linear predictor through a monotonic differentiable link function $g$.

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

Here, $\mathbf{x}_i$ is a fixed known vector of explanatory variables, and $\boldsymbol{\beta}$ is a vector of unknown parameters.

### Log-Likelihood Functions

Log-likelihood functions for the distributions that are available in the procedure are parameterized in terms of the means $\mu_i$ and the dispersion parameter $\phi$. The term $y_i$ represents the response for the $i$th observation, and $w_i$ represents the known dispersion weight. The log-likelihood functions are of the form

$$L(\mathbf{y}, \boldsymbol{\mu}, \phi) = \sum_i \log\left(f(y_i, \mu_i, \phi)\right)$$

where the sum is over the observations. The forms of the individual contributions

$$l_i = \log\left(f(y_i, \mu_i, \phi)\right)$$

are shown in the following list; the parameterizations are expressed in terms of the mean and dispersion parameters.

- Normal:

$$l_i = -\frac{1}{2}\left[\frac{w_i(y_i-\mu_i)^2}{\phi} + \log\left(\frac{\phi}{w_i}\right) + \log(2\pi)\right]$$

- Inverse Gaussian:

$$l_i = -\frac{1}{2}\left[\frac{w_i(y_i-\mu_i)^2}{y_i\mu^2\phi} + \log\left(\frac{\phi y_i^3}{w_i}\right) + \log(2\pi)\right]$$

- Gamma:

$$l_i = \frac{w_i}{\phi}\log\left(\frac{w_i y_i}{\phi\mu_i}\right) - \frac{w_i y_i}{\phi\mu_i} - \log(y_i) - \log\left(\Gamma\left(\frac{w_i}{\phi}\right)\right)$$

- Negative Binomial:

$$l_i = y\log(k\mu) - (y + 1/k)\log(1 + k\mu) + \log\left(\frac{\Gamma(y + 1/k)}{\Gamma(y + 1)\Gamma(1/k)}\right)$$

- Poisson:

$$l_i = y_i\log(\mu_i) - \mu_i$$

- Binomial:

$$l_i = [r_i\log(p_i) + (n_i - r_i)\log(1 - p_i)]$$

- Multinomial:

$$l_i = \sum_j y_{ij}\log(\mu_{ij})$$

For the binomial, multinomial, and Poisson distribution, terms involving binomial coefficients or factorials of the observed counts are dropped from the computation of the log-likelihood function since they do not affect parameter estimates or their estimated covariances.

### Maximum Likelihood Fitting

The GENMOD procedure uses a ridge-stabilized Newton-Raphson algorithm to maximize the log-likelihood function $L(\mathbf{y}, \boldsymbol{\mu}, \phi)$ with respect to the regression parameters. By default, the procedure also produces maximum likelihood estimates of the scale parameter as defined in the "Response Probability Distributions" section (page 1402) for the normal, inverse Gaussian, negative binomial, and gamma distributions.

On the $r$th iteration, the algorithm updates the parameter vector $\boldsymbol{\beta}_r$ with

$$\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r - \mathbf{H}^{-1}\mathbf{s}$$

where $\mathbf{H}$ is the Hessian (second derivative) matrix, and $\mathbf{s}$ is the gradient (first derivative) vector of the log-likelihood function, both evaluated at the current value of the parameter vector. That is,

$$\mathbf{s} = [s_j] = \left[\frac{\partial L}{\partial\beta_j}\right]$$

and

$$\mathbf{H} = [h_{ij}] = \left[\frac{\partial^2 L}{\partial\beta_i\partial\beta_j}\right]$$

In some cases, the scale parameter is estimated by maximum likelihood. In these cases, elements corresponding to the scale parameter are computed and included in $\mathbf{s}$ and $\mathbf{H}$.

If $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ is the linear predictor for observation $i$ and $g$ is the link function, then $\eta_i = g(\mu_i)$, so that $\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$ is an estimate of the mean of the $i$th observation, obtained from an estimate of the parameter vector $\boldsymbol{\beta}$.

The gradient vector and Hessian matrix for the regression parameters are given by

$$\begin{aligned} \mathbf{s} &= \sum_i \frac{w_i(y_i - \mu_i)\mathbf{x}_i}{V(\mu_i)g'(\mu_i)\phi} \\ \mathbf{H} &= -\mathbf{X}'\mathbf{W}_o\mathbf{X} \end{aligned}$$

where $\mathbf{X}$ is the design matrix, $\mathbf{x}_i$ is the transpose of the $i$th row of $\mathbf{X}$, and $V$ is the variance function. The matrix $\mathbf{W}_o$ is diagonal with its $i$th diagonal element

$$w_{oi} = w_{ei} + w_i(y_i - \mu_i)\frac{V(\mu_i)g''(\mu_i) + V'(\mu_i)g'(\mu_i)}{(V(\mu_i))^2(g'(\mu_i))^3\phi}$$

where

$$w_{ei} = \frac{w_i}{\phi V(\mu_i)(g'(\mu_i))^2}$$

The primes denote derivatives of $g$ and $V$ with respect to $\mu$. The negative of $\mathbf{H}$ is called the observed information matrix. The expected value of $\mathbf{W}_o$ is a diagonal matrix $\mathbf{W}_e$ with diagonal values $w_{ei}$. If you replace $\mathbf{W}_o$ with $\mathbf{W}_e$, then the negative of $\mathbf{H}$ is called the expected information matrix. $\mathbf{W}_e$ is the weight matrix for the Fisher's scoring method of fitting. Either $\mathbf{W}_o$ or $\mathbf{W}_e$ can be used in the update equation. The GENMOD procedure uses Fisher's scoring for iterations up to the number specified by the SCORING option in the MODEL statement, and it uses the observed information matrix on additional iterations.

### Covariance and Correlation Matrix

The estimated covariance matrix of the parameter estimator is given by

$$\boldsymbol{\Sigma} = -\mathbf{H}^{-1}$$

where $\mathbf{H}$ is the Hessian matrix evaluated using the parameter estimates on the last iteration. Note that the dispersion parameter, whether estimated or specified, is incorporated into $\mathbf{H}$. Rows and columns corresponding to aliased parameters are not included in $\boldsymbol{\Sigma}$.

The correlation matrix is the normalized covariance matrix. That is, if $\sigma_{ij}$ is an element of $\boldsymbol{\Sigma}$, then the corresponding element of the correlation matrix is $\sigma_{ij}/\sigma_i\sigma_j$, where $\sigma_i = \sqrt{\sigma_{ii}}$.

### Goodness of Fit

Two statistics that are helpful in assessing the goodness of fit of a given generalized linear model are the scaled deviance and Pearson's chi-square statistic. For a fixed value of the dispersion parameter $\phi$, the scaled deviance is defined to be twice the difference between the maximum achievable log likelihood and the log likelihood at the maximum likelihood estimates of the regression parameters.

Note that these statistics are not valid for GEE models.

If $l(\mathbf{y}, \boldsymbol{\mu})$ is the log-likelihood function expressed as a function of the predicted mean values $\boldsymbol{\mu}$ and the vector $\mathbf{y}$ of response values, then the scaled deviance is defined by

$$D^*(\mathbf{y}, \boldsymbol{\mu}) = 2(l(\mathbf{y}, \mathbf{y}) - l(\mathbf{y}, \boldsymbol{\mu}))$$

For specific distributions, this can be expressed as

$$D^*(\mathbf{y}, \boldsymbol{\mu}) = \frac{D(\mathbf{y}, \boldsymbol{\mu})}{\phi}$$

where $D$ is the deviance. The following table displays the deviance for each of the probability distributions available in PROC GENMOD.

| Distribution | Deviance |
|---|---|
| normal | $\sum_i w_i(y_i - \mu_i)^2$ |
| Poisson | $2\sum_i w_i \left[ y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i)\right]$ |
| binomial | $2\sum_i w_i m_i \left[ y_i \log\left(\frac{y_i}{\mu_i}\right) + (1 - y_i) \log\left(\frac{1-y_i}{1-\mu_i}\right)\right]$ |
| gamma | $2\sum_i w_i \left[ -\log\left(\frac{y_i}{\mu_i}\right) + \frac{y_i - \mu_i}{\mu_i}\right]$ |
| inverse Gaussian | $\sum_i \frac{w_i(y_i-\mu_i)^2}{\mu_i^2 y_i}$ |
| multinomial | $\sum_i \sum_j w_i y_{ij} \log\left(\frac{y_{ij}}{p_{ij} m_i}\right)$ |
| negative binomial | $2\sum_i w_i \left[ y \log(y/\mu) - (y + 1/k) \log\left(\frac{y+1/k}{\mu+1/k}\right)\right]$ |

In the binomial case, $y_i = r_i/m_i$, where $r_i$ is a binomial count and $m_i$ is the binomial number of trials parameter.

In the multinomial case, $y_{ij}$ refers to the observed number of occurrences of the $j$th category for the $i$th subpopulation defined by the AGGREGATE= variable, $m_i$ is the total number in the $i$th subpopulation, and $p_{ij}$ is the category probability.

Pearson's chi-square statistic is defined as

$$X^2 = \sum_i \frac{w_i(y_i - \mu_i)^2}{V(\mu_i)}$$

and the scaled Pearson's chi-square is $X^2/\phi$.

The scaled version of both of these statistics, under certain regularity conditions, has a limiting chi-square distribution, with degrees of freedom equal to the number of observations minus the number of parameters estimated. The scaled version can be used as an approximate guide to the goodness of fit of a given model. Use caution before applying these statistics to ensure that all the conditions for the asymptotic distributions hold. McCullagh and Nelder (1989) advise that differences in deviances for nested models can be better approximated by chi-square distributions than the deviances themselves.

In cases where the dispersion parameter is not known, an estimate can be used to obtain an approximation to the scaled deviance and Pearson's chi-square statistic. One strategy is to fit a model that contains a sufficient number of parameters so that all systematic variation is removed, estimate $\phi$ from this model, and then use this estimate in computing the scaled deviance of sub-models. The deviance or Pearson's chi-square divided by its degrees of freedom is sometimes used as an estimate of the dispersion parameter $\phi$. For example, since the limiting chi-square distribution of the scaled deviance $D^* = D/\phi$ has $n - p$ degrees of freedom, where $n$ is the number of observations and $p$ the number of parameters, equating $D^*$ to its mean and solving for $\phi$ yields $\hat{\phi} = D/(n-p)$. Similarly, an estimate of $\phi$ based on Pearson's chi-square $X^2$ is $\hat{\phi} = X^2/(n-p)$. Alternatively, a maximum likelihood estimate of $\phi$ can be computed by the procedure, if desired. See the discussion in the "Type 1 Analysis" section on page 1413 for more on the estimation of the dispersion parameter.

### Dispersion Parameter

There are several options available in PROC GENMOD for handling the exponential distribution dispersion parameter. The NOSCALE and SCALE options in the MODEL statement affect the way in which the dispersion parameter is treated. If you specify the SCALE=DEVIANCE option, the dispersion parameter is estimated by the deviance divided by its degrees of freedom. If you specify the SCALE=PEARSON option, the dispersion parameter is estimated by Pearson's chi-square statistic divided by its degrees of freedom.

Otherwise, values of the SCALE and NOSCALE options and the resultant actions are displayed in the following table.

| **NOSCALE** | **SCALE=*value*** | **Action** |
|---|---|---|
| present | present | scale fixed at *value* |
| present | not present | scale fixed at 1 |
| not present | not present | scale estimated by ML |
| not present | present | scale estimated by ML, starting point at *value* |

The meaning of the scale parameter displayed in the "Analysis Of Parameter Estimates" table is different for the Gamma distribution than for the other distributions. The relation of the scale parameter as used by PROC GENMOD to the exponential family dispersion parameter $\phi$ is displayed in the following table. For the binomial and Poisson distributions, $\phi$ is the overdispersion parameter, as defined in the "Overdispersion" section, which follows.

| Distribution | Scale |
|---|---|
| normal | $\sqrt{\phi}$ |
| inverse Gaussian | $\sqrt{\phi}$ |
| gamma | $1/\phi$ |
| binomial | $\sqrt{\phi}$ |
| Poisson | $\sqrt{\phi}$ |

In the case of the negative binomial distribution, PROC GENMOD reports the "dispersion" parameter estimated by maximum likelihood. This is the negative binomial parameter $k$ defined in the "Response Probability Distributions" section (page 1402).

### Overdispersion

Overdispersion is a phenomenon that sometimes occurs in data that are modeled with the binomial or Poisson distributions. If the estimate of dispersion after fitting, as measured by the deviance or Pearson's chi-square, divided by the degrees of freedom, is not near 1, then the data may be *overdispersed* if the dispersion estimate is greater than 1 or *underdispersed* if the dispersion estimate is less than 1. A simple way to model this situation is to allow the variance functions of these distributions to have a multiplicative overdispersion factor $\phi$.

- binomial : $V(\mu) = \phi\mu(1-\mu)$
- Poisson : $V(\mu) = \phi\mu$

The models are fit in the usual way, and the parameter estimates are not affected by the value of $\phi$. The covariance matrix, however, is multiplied by $\phi$, and the scaled deviance and log likelihoods used in likelihood ratio tests are divided by $\phi$. The profile likelihood function used in computing confidence intervals is also divided by $\phi$. If you specify an WEIGHT statement, $\phi$ is divided by the value of the WEIGHT variable for each observation. This has the effect of multiplying the contributions of the log-likelihood function, the gradient, and the Hessian by the value of the WEIGHT variable for each observation.

The SCALE= option in the MODEL statement enables you to specify a value of $\sigma = \sqrt{\phi}$ for the binomial and Poisson distributions. If you specify the SCALE=DEVIANCE option in the MODEL statement, the procedure uses the deviance divided by degrees of freedom as an estimate of $\phi$, and all statistics are adjusted appropriately. You can use Pearson's chi-square instead of the deviance by specifying the SCALE=PEARSON option.

The function obtained by dividing a log-likelihood function for the binomial or Poisson distribution by a dispersion parameter is not a legitimate log-likelihood function. It is an example of a *quasi-likelihood* function. Most of the asymptotic theory for log likelihoods also applies to quasi-likelihoods, which justifies computing standard errors and likelihood ratio statistics using quasi-likelihoods instead of proper log likelihoods. Refer to McCullagh and Nelder (1989, Chapter 9) and McCullagh (1983) for details on quasi-likelihood functions.

Although the estimate of the dispersion parameter is often used to indicate overdispersion or underdispersion, this estimate may also indicate other problems such as an incorrectly specified model or outliers in the data. You should carefully assess whether this type of model is appropriate for your data.

## Specification of Effects

Each term in a model is called an effect. Effects are specified in the MODEL statement in the same way as in the GLM procedure. You specify effects with a special notation that uses variable names and operators. There are two types of variables, *classification* (or *class*) variables and *continuous* variables. There are two primary types of operators, *crossing* and *nesting*. A third type, the *bar* operator, is used to simplify effect specification. Crossing is the type of operator most commonly used in generalized linear models.

Variables that identify classification levels are called *class* variables in the SAS System and are identified in a CLASS statement. These may also be called *categorical, qualitative, discrete,* or *nominal* variables. Class variables can be either character or numeric. The values of class variables are called *levels*. For example, the class variable Sex could have levels ‘male’ and ‘female’.

In a model, an explanatory variable that is not declared in a CLASS statement is assumed to be continuous. Continuous variables must be numeric. For example, the heights and weights of subjects in an experiment are continuous variables.

The types of effects most useful in generalized linear models are shown in the following list. Assume that A, B, and C are class variables and that X1 and X2 are continuous variables.

- Regressor effects are specified by writing continuous variables by themselves: X1, X2.
- Polynomial effects are specified by joining two or more continuous variables with asterisks: X1*X2.
- Main effects are specified by writing class variables by themselves: A, B, C.
- Crossed effects (interactions) are specified by joining two or more class variables with asterisks: A*B, B*C, A*B*C.
- Nested effects are specified by following a main effect or crossed effect with a class variable or list of class variables enclosed in parentheses: B(A), C(B A), A*B(C). In the preceding example, B(A) is “B nested within A.”
- Combinations of continuous and class variables can be specified in the same way using the crossing and nesting operators.

The bar operator consists of two effects joined with a vertical bar (|). It is shorthand notation for including the left-hand side, the right-hand side, and the cross between them as effects in the model. For example, A | B is equivalent to A B A*B. The effects in the bar operator can be class variables, continuous variables, or combinations of effects defined using operators. Multiple bars are permitted. For example, A | B | C means A B C A*B A*C B*C A*B*C.

You can specify the maximum number of variables in any effect that results from bar evaluation by specifying the maximum number, preceded by an @ sign. For example, A | B | C@2 results in effects that involve two or fewer variables: A B C A*B A*C B*C.

For further information on types of effects and their specification, see Chapter 30, "The GLM Procedure."

## Parameterization Used in PROC GENMOD

### Design Matrix

The linear predictor part of a generalized linear model is

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is an unknown parameter vector and $\mathbf{X}$ is a known design matrix. By default, all models automatically contain an intercept term; that is, the first column of $\mathbf{X}$ contains all 1s. Additional columns of $\mathbf{X}$ are generated for classification variables, regression variables, and any interaction terms included in the model. PROC GENMOD parameterizes main effects and interaction terms using the same ordering rules that PROC GLM uses. This is important to understand when you want to construct likelihood ratios for custom contrasts using the CONTRAST statement. See Chapter 30, "The GLM Procedure," for more details on model parameterization.

Some columns of $\mathbf{X}$ can be linearly dependent on other columns due to specifying an overparameterized model. For example, when you specify a model consisting of an intercept term and a class variable, the column corresponding to any one of the levels of the class variable is linearly dependent on the other columns of $\mathbf{X}$. PROC GENMOD handles this in the same manner as PROC GLM. The columns of $\mathbf{X}'\mathbf{X}$ are checked in the order in which the model is specified for dependence on preceding columns. If a dependency is found, the parameter corresponding to the dependent column is set to 0 along with its standard error to indicate that it is not estimated. The order in which the levels of a class variable are checked for dependencies can be set by the ORDER= option in the PROC GENMOD statement.

You can exclude the intercept term from the model by specifying the NOINT option in the MODEL statement.

### *Missing Level Combinations*

All levels of interaction terms involving classification variables may not be represented in the data. In that case, PROC GENMOD does not include parameters in the model for the missing levels.

## Type 1 Analysis

A Type 1 analysis consists of fitting a sequence of models, beginning with a simple model with only an intercept term, and continuing through a model of specified complexity, fitting one additional effect on each step. Likelihood ratio statistics, that is, twice the difference of the log likelihoods, are computed between successive models. This type of analysis is sometimes called an analysis of deviance since, if the dispersion parameter is held fixed for all models, it is equivalent to computing differences of scaled deviances. The asymptotic distribution of the likelihood ratio statistics, under the hypothesis that the additional parameters included in the model are equal to 0, is a chi-square with degrees of freedom equal to the difference in the number of parameters estimated in the successive models. Thus, these statistics can be used in a test of hypothesis of the significance of each additional term fit.

This type of analysis is not available for GEE models, since the deviance is not computed for this type of model.

If the dispersion parameter $\phi$ is known, it can be included in the models; if it is unknown, there are two strategies allowed by PROC GENMOD. The dispersion parameter can be estimated from a maximal model by the deviance or Pearson's chi-square divided by degrees of freedom, as discussed in the "Goodness of Fit" section on page 1408, and this value can be used in all models. An alternative is to consider the dispersion to be an additional unknown parameter for each model and estimate it by maximum likelihood on each step. By default, PROC GENMOD estimates scale by maximum likelihood at each step.

A table of likelihood ratio statistics is produced, along with associated $p$-values based on the asymptotic chi-square distributions.

If you specify either the SCALE=DEVIANCE or the SCALE=PEARSON option in the MODEL statement, the dispersion parameter is estimated using the deviance or Pearson's chi-square statistic, and $F$ statistics are computed in addition to the chi-square statistics for assessing the significance of each additional term in the Type 1 analysis. See the section "F Statistics" on page 1416 for a definition of $F$ statistics.

This Type 1 analysis has the general property that the results depend on the order in which the terms of the model are fitted. The terms are fitted in the order in which they are specified in the MODEL statement.

## Type 3 Analysis

A Type 3 analysis is similar to the Type III sums of squares used in PROC GLM, except that likelihood ratios are used instead of sums of squares. First, a Type III estimable function is defined for an effect of interest in exactly the same way as in PROC GLM. Then, maximum likelihood estimation is performed under the constraint that the Type III function of the parameters is equal to 0, using constrained optimization. Let the resulting constrained parameter estimates be $\tilde{\boldsymbol{\beta}}$ and the log likelihood be $l(\tilde{\boldsymbol{\beta}})$. Then the likelihood ratio statistic

$$S = 2(l(\hat{\boldsymbol{\beta}}) - l(\tilde{\boldsymbol{\beta}}))$$

where $\hat{\boldsymbol{\beta}}$ is the unconstrained estimate, has an asymptotic chi-square distribution under the hypothesis that the Type III contrast is equal to 0, with degrees of freedom equal to the number of parameters associated with the effect.

When a Type 3 analysis is requested, PROC GENMOD produces a table that contains the likelihood ratio statistics, degrees of freedom, and $p$-values based on the limiting chi-square distributions for each effect in the model. If you specify either the DSCALE or PSCALE option in the MODEL statement, $F$ statistics are also computed for each effect.

Options for handling the dispersion parameter are the same as for a Type 1 analysis. The dispersion parameter can be specified to be a known value, estimated from the deviance or Pearson's chi-square divided by degrees of freedom, or estimated by maximum likelihood individually for the unconstrained and constrained models. By default, PROC GENMOD estimates scale by maximum likelihood for each model fit.

The results of this type of analysis do not depend on the order in which the terms are specified in the MODEL statement.

A Type 3 analysis can consume considerable computation time since a constrained model is fitted for each effect. Wald statistics for Type 3 contrasts are computed if you specify the WALD option. Wald statistics for contrasts use less computation time than likelihood ratio statistics but may be less accurate indicators of the significance of the effect of interest. The Wald statistic for testing $\mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{L}$ is the contrast matrix, is defined by

$$S = (\mathbf{L}'\hat{\boldsymbol{\beta}})'(\mathbf{L}'\hat{\boldsymbol{\Sigma}}\mathbf{L})^{-}(\mathbf{L}'\hat{\boldsymbol{\beta}})$$

where $\boldsymbol{\beta}$ is the maximum likelihood estimate and $\boldsymbol{\Sigma}$ is its estimated covariance matrix. The asymptotic distribution of $S$ is chi-square with $r$ degrees of freedom, where $r$ is the rank of $\mathbf{L}$.

See Chapter 30, "The GLM Procedure,"and Chapter 12, "The Four Types of Estimable Functions," for more information about Type III estimable functions. Also refer to Littell, Freund, and Spector (1991).

Generalized score tests for Type III contrasts are computed for GEE models if you specify the TYPE3 option in the MODEL statement when a REPEATED statement

is also used. See the section "Generalized Score Statistics" on page 1428 for more information on generalized score statistics. Wald tests are also available with the Wald option in the CONTRAST statement.

## Confidence Intervals for Parameters

### Likelihood Ratio-Based Confidence Intervals

PROC GENMOD produces likelihood ratio-based confidence intervals, also known as profile likelihood confidence intervals, for parameter estimates for generalized linear models. These are not computed for GEE models, since there is no likelihood for this type of model. Suppose that the parameter vector is $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]'$ and that you want a confidence interval for $\beta_j$. The profile likelihood function for $\beta_j$ is defined as

$$l^*(\beta_j) = \max_{\tilde{\boldsymbol{\beta}}} l(\boldsymbol{\beta})$$

where $\tilde{\boldsymbol{\beta}}$ is the vector $\boldsymbol{\beta}$ with the $j$th element fixed at $\beta_j$ and $l$ is the log likelihood function. If $l = l(\hat{\boldsymbol{\beta}})$ is the log likelihood evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$, then $2(l - l^*(\beta_j))$ has a limiting chi-square distribution with one degree of freedom if $\beta_j$ is the true parameter value. A $(1-\alpha)100\%$ confidence interval for $\beta_j$ is

$$\left\{\beta_j : l^*(\beta_j) \geq l_0 = l - 0.5\chi^2_{1-\alpha,1}\right\}$$

where $\chi^2_{1-\alpha,1}$ is the $100(1-\alpha)$ percentile of the chi-square distribution with one degree of freedom. The endpoints of the confidence interval can be found by solving numerically for values of $\beta_j$ that satisfy equality in the preceding relation. PROC GENMOD solves this by starting at the maximum likelihood estimate of $\boldsymbol{\beta}$. The log likelihood function is approximated with a quadratic surface, for which an exact solution is possible. The process is iterated until convergence to an endpoint is attained. The process is repeated for the other endpoint.

Convergence is controlled by the CICONV= option in the MODEL statement. Suppose $\epsilon$ is the number specified in the CICONV= option. The default value of $\epsilon$ is $10^{-4}$. Let the parameter of interest be $\beta_j$ and define $\mathbf{r} = \mathbf{u}_j$, the unit vector with a 1 in position $j$ and 0s elsewhere. Convergence is declared on the current iteration if the following two conditions are satisfied:

$$\begin{aligned} |l^*(\beta_j) - l_0| &\leq \epsilon \\ (\mathbf{s} + \lambda\mathbf{r})'\mathbf{H}^{-1}(\mathbf{s} + \lambda\mathbf{r}) &\leq \epsilon \end{aligned}$$

where $l^*(\beta_j)$, $\mathbf{s}$, and $\mathbf{H}$ are the log likelihood, the gradient, and the Hessian evaluated at the current parameter vector and $\lambda$ is a constant computed by the procedure. The first condition for convergence means that the log-likelihood function must be within $\epsilon$ of the correct value, and the second condition means that the gradient vector must be proportional to the restriction vector $\mathbf{r}$.

When you request the LRCI option in the MODEL statement, PROC GENMOD computes profile likelihood confidence intervals for all parameters in the model, including the scale parameter, if there is one. The interval endpoints are displayed in a table as well as the values of the remaining parameters at the solution.

### Wald Confidence Intervals

You can request that PROC GENMOD produce Wald confidence intervals for the parameters. The (1-$\alpha$)100% Wald confidence interval for a parameter $\beta$ is defined as

$$\hat{\beta} \pm z_{1-\alpha/2}\hat{\sigma}$$

where $z_p$ is the $100p$th percentile of the standard normal distribution, $\hat{\beta}$ is the parameter estimate, and $\hat{\sigma}$ is the estimate of its standard error.

## F Statistics

Suppose that $D_0$ is the deviance resulting from fitting a generalized linear model and that $D_1$ is the deviance from fitting a submodel. Then, under appropriate regularity conditions, the asymptotic distribution of $(D_1 - D_0)/\phi$ is chi-square with $r$ degrees of freedom, where $r$ is the difference in the number of parameters between the two models and $\phi$ is the dispersion parameter. If $\phi$ is unknown, and $\hat{\phi}$ is an estimate of $\phi$ based on the deviance or Pearson's chi-square divided by degrees of freedom, then, under regularity conditions, $(n-p)\hat{\phi}/\phi$ has an asymptotic chi-square distribution with $n-p$ degrees of freedom. Here, $n$ is the number of observations and $p$ is the number of parameters in the model that is used to estimate $\phi$. Thus, the asymptotic distribution of

$$F = \frac{D_1 - D_0}{r\hat{\phi}}$$

is the $F$ distribution with $r$ and $n-p$ degrees of freedom, assuming that $(D_1 - D_0)/\phi$ and $(n-p)\hat{\phi}/\phi$ are approximately independent.

This $F$ statistic is computed for the Type 1 analysis, Type 3 analysis, and hypothesis tests specified in CONTRAST statements when the dispersion parameter is estimated by either the deviance or Pearson's chi-square divided by degrees of freedom, as specified by the DSCALE or PSCALE option in the MODEL statement. In the case of a Type 1 analysis, model 0 is the higher-order model obtained by including one additional effect in model 1. For a Type 3 analysis and hypothesis tests, model 0 is the full specified model and model 1 is the sub-model obtained from constraining the Type III contrast or the user-specified contrast to be 0.

## Lagrange Multiplier Statistics

When you select the NOINT or NOSCALE option, restrictions are placed on the intercept or scale parameters. Lagrange multiplier, or score, statistics are computed in these cases. These statistics assess the validity of the restrictions, and they are computed as

$$\chi^2 = \frac{s^2}{V}$$

where $s$ is the component of the score vector evaluated at the restricted maximum corresponding to the restricted parameter and $V = \mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21}$. The matrix $\mathbf{I}$ is the information matrix, 1 refers to the restricted parameter, and 2 refers to the rest of the parameters.

Under regularity conditions, this statistic has an asymptotic chi-square distribution with one degree of freedom, and $p$-values are computed based on this limiting distribution.

Refer to Rao (1973, p. 417) for details.

## Predicted Values of the Mean

### Predicted Values

A predicted value, or fitted value, of the mean $\mu_i$ corresponding to the vector of covariates $\mathbf{x}_i$ is given by

$$\hat{\mu}_i = g^{-1}(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$$

where $g$ is the link function, regardless of whether $\mathbf{x}_i$ corresponds to an observation or not. That is, the response variable can be missing and the predicted value is still computed for valid $\mathbf{x}_i$. In the case where $\mathbf{x}_i$ does not correspond to a valid observation, $\mathbf{x}_i$ is not checked for estimability. You should check the estimability of $\mathbf{x}_i$ in this case in order to ensure the uniqueness of the predicted value of the mean. If there is an offset, it is included in the predicted value computation.

### Confidence Intervals on Predicted Values

Approximate confidence intervals for predicted values of the mean can be computed as follows. The variance of the linear predictor $\eta_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ is estimated by

$$\sigma_x^2 = \mathbf{x}_i'\boldsymbol{\Sigma}\mathbf{x}_i$$

where $\boldsymbol{\Sigma}$ is the estimated covariance of $\hat{\boldsymbol{\beta}}$.

Approximate $100(1-\alpha)\%$ confidence intervals are computed as

$$g^{-1}\left(\mathbf{x}_i'\hat{\beta} \pm z_{1-\alpha/2}\sigma_x\right)$$

where $z_p$ is the $100p$ percentile of the standard normal distribution and $g$ is the link function. If either endpoint in the argument is outside the valid range of arguments for the inverse link function, the corresponding confidence interval endpoint is set to missing.

## Residuals

The GENMOD procedure computes three kinds of residuals. The raw residual is defined as

$$r_i = y_i - \mu_i$$

where $y_i$ is the $i$th response and $\mu_i$ is the corresponding predicted mean.

The Pearson residual is the square root of the $i$th contribution to the Pearson's chi-square.

$$r_{Pi} = (y_i - \mu_i)\sqrt{\frac{w_i}{V(\mu_i)}}$$

Finally, the deviance residual is defined as the square root of the contribution of the $i$th observation to the deviance, with the sign of the raw residual.

$$r_{Di} = \sqrt{d_i}(\mathit{sign}(y_i - \mu_i))$$

The adjusted Pearson, deviance, and likelihood residuals are defined by Agresti (1990), Williams (1987), and Davison and Snell (1991). These residuals are useful for outlier detection and for assessing the influence of single observations on the fitted model.

For the generalized linear model, the variance of the $i$th individual observation is given by

$$v_i = \frac{\phi V(\mu_i)}{w_i}$$

where $\phi$ is the dispersion parameter, $w_i$ is a user-specified prior weight (if not specified, $w_i = 1$), $\mu_i$ is the mean, and $V(\mu_i)$ is the variance function. Let

$$w_{ei} = v_i^{-1}(g'(\mu_i))^{-2}$$

for the $i$th observation, where $g'(\mu_i)$ is the derivative of the link function, evaluated at $\mu_i$. Let $\mathbf{W}_e$ be the diagonal matrix with $w_{ei}$ denoting the $i$th diagonal element. The weight matrix $\mathbf{W}_e$ is used in computing the expected information matrix.

Define $h_i$ as the $i$th diagonal element of the matrix

$$\mathbf{W}_e^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}_e\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_e^{\frac{1}{2}}$$

The Pearson residuals, standardized to have unit asymptotic variance, are given by

$$r_{Pi} = \frac{y_i - \mu_i}{\sqrt{v_i(1-h_i)}}$$

The deviance residuals, standardized to have unit asymptotic variance, are given by

$$r_{Di} = \frac{sign(y_i - \mu_i)\sqrt{d_i}}{\sqrt{\phi(1-h_i)}}$$

where $d_i$ is the square root of the contribution to the total deviance from observation $i$, and sign$(y_i - \mu_i)$ is 1 if $y_i - \mu_i$ is positive and -1 if $y_i - \mu_i$ is negative. The likelihood residuals are defined by

$$r_{Gi} = sign(y_i - \mu_i)\sqrt{(1-h_i)r_{Di}^2 + h_i r_{Pi}^2}$$

## Multinomial Models

This type of model applies to cases where an observation can fall into one of $k$ categories. Binary data occurs in the special case where $k = 2$. If there are $m_i$ observations in a subpopulation $i$, then the probability distribution of the number falling into the $k$ categories $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots y_{ik})$ can be modeled by the multinomial distribution, defined in the "Response Probability Distributions" section (page 1402), with $\sum_j y_{ij} = m_i$. The multinomial model is an *ordinal* model if the categories have a natural order.

The GENMOD procedure orders the response categories for ordinal multinomial models from lowest to highest by default. This is different from the binomial distribution, where the response probability for the highest of the two categories is modeled. You can change the way GENMOD orders the response levels with the RORDER= option in the PROC GENMOD statement. The order that GENMOD uses is shown in the "Response Profiles" output table described in the section "Response Profile" on page 1429.

The GENMOD procedure supports only the ordinal multinomial model. If $(p_{i1}, p_{i2}, \cdots p_{ik})$ are the category probabilities, the cumulative category probabilities are modeled with the same link functions used for binomial data. Let $P_{ir} = \sum_{j=1}^{r} p_{ij}$, $r = 1, 2, \cdots, k-1$ be the cumulative category probabilities (note that $P_{ik} = 1$). The ordinal model is

$$g(P_{ir}) = \mu_r + \mathbf{x}_i{}'\boldsymbol{\beta} \quad \text{for} \quad r = 1, 2, \cdots k-1$$

where $\mu_1, \mu_2, \cdots \mu_{k-1}$ are intercept terms that depend only on the categories and $\mathbf{x}_i$ is a vector of covariates that does not include an intercept term. The logit, probit, and

complementary log-log link functions $g$ are available. These are obtained by specifying the MODEL statement options DIST=MULTINOMIAL and LINK=CUMLOGIT (cumulative logit), LINK=CUMPROBIT (cumulative probit), or LINK=CUMCLL (cumulative complementary log-log). Alternatively,

$$P_{ir} = \mathrm{F}(\mu_r + \mathbf{x}_i{}'\boldsymbol{\beta}) \quad \text{for} \quad r = 1, 2, \cdots k-1$$

where $\mathrm{F} = g^{-1}$ is a cumulative distribution function for the logistic, normal, or extreme value distribution.

PROC GENMOD estimates the intercept parameters $\mu_1, \mu_2, \cdots \mu_{k-1}$ and regression parameters $\boldsymbol{\beta}$ by maximum likelihood.

The subpopulations $i$ are defined by constant values of the AGGREGATE= variable. This has no effect on the parameter estimates, but it does affect the deviance and Pearson chi-square statistics; it also affects parameter estimate standard errors if you specify the SCALE=DEVIANCE or SCALE=PEARSON options.

## Generalized Estimating Equations

Let $Y_{ij}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, K$ represent the $j$th measurement on the $i$th subject. There are $n_i$ measurements on subject $i$ and $\sum_{i=1}^{K} n_i$ total measurements.

Correlated data are modeled using the same link function and linear predictor setup (systematic component) as the independence case. The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated measurements must also be modeled. Let the vector of measurements on the $i$th subject be $\mathbf{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]'$ with corresponding vector of means $\boldsymbol{\mu}_i = [\mu_{i1}, \ldots, \mu_{in_i}]'$ and let $\mathbf{V}_i$ be the covariance matrix of $\mathbf{Y}_i$. Let the vector of independent, or explanatory, variables for the $j$th measurement on the $i$th subject be

$$\mathbf{X}_{ij} = [x_{ij1}, \ldots, x_{ijp}]'$$

The Generalized Estimating Equation of Liang and Zeger (1986) for estimating the $p \times 1$ vector of regression parameters $\boldsymbol{\beta}$ is an extension of the independence estimating equation to correlated data and is given by

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}$$

Since

$$g(\mu_{ij}) = \mathbf{x}_{ij}{}'\boldsymbol{\beta}$$

where $g$ is the link function, the $p \times n_i$ matrix of partial derivatives of the mean with respect to the regression parameters for the $i$th subject is given by

$$\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{x_{i11}}{g'(\mu_{i1})} & \cdots & \frac{x_{in_i1}}{g'(\mu_{in_i})} \\ \vdots & & \vdots \\ \frac{x_{i1p}}{g'(\mu_{i1})} & \cdots & \frac{x_{in_ip}}{g'(\mu_{in_i})} \end{bmatrix}$$

### *Working Correlation Matrix*

Let $\mathbf{R}_i(\boldsymbol{\alpha})$ be an $n_i \times n_i$ "working" correlation matrix that is fully specified by the vector of parameters $\boldsymbol{\alpha}$. The covariance matrix of $\mathbf{Y}_i$ is modeled as

$$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{\frac{1}{2}}$$

where $\mathbf{A}_i$ is an $n_i \times n_i$ diagonal matrix with $v(\mu_{ij})$ as the $j$th diagonal element. If $\mathbf{R}_i(\boldsymbol{\alpha})$ is the true correlation matrix of $\mathbf{Y}_i$, then $\mathbf{V}_i$ is the true covariance matrix of $\mathbf{Y}_i$.

The working correlation matrix is usually unknown and must be estimated. It is estimated in the iterative fitting process using the current value of the parameter vector $\boldsymbol{\beta}$ to compute appropriate functions of the Pearson residual

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$$

If you specify the working correlation as $\mathbf{R}_0 = \mathbf{I}$, which is the identity matrix, the GEE reduces to the independence estimating equation.

Following are the structures of the working correlation supported by the GENMOD procedure and the estimators used to estimate the working correlations.

| Working Correlation Structure | Estimator |
|---|---|
| **Fixed**<br>$\text{Corr}(Y_{ij}, Y_{ik}) = r_{jk}$<br>where $r_{jk}$ is the $jk$th element of a constant, user-specified correlation matrix $\mathbf{R}_0$. | The working correlation is not estimated in this case. |
| **Independent**<br>$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}$ | The working correlation is not estimated in this case. |
| ***m*-dependent**<br>$\text{Corr}(Y_{ij}, Y_{i,j+t}) = \begin{cases} 1 & t = 0 \\ \alpha_t & t = 1, 2, \ldots, m \\ 0 & t > m \end{cases}$ | $\hat{\alpha}_t = \frac{1}{(K_t - p)\phi} \sum_{i=1}^{K} \sum_{j \leq n_i - t} e_{ij} e_{i,j+t}$<br>$K_t = \sum_{i=1}^{K} (n_i - t)$ |
| **Exchangeable**<br>$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha & j \neq k \end{cases}$ | $\hat{\alpha} = \frac{1}{(N^* - p)\phi} \sum_{i=1}^{K} \sum_{j \neq k} e_{ij} e_{ik}$<br>$N^* = \sum_{i=1}^{K} n_i (n_i - 1)$ |
| **Unstructured**<br>$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha_{jk} & j \neq k \end{cases}$ | $\hat{\alpha}_{jk} = \frac{1}{(K - p)\phi} \sum_{i=1}^{K} e_{ij} e_{ik}$ |
| **Autoregressive AR(1)**<br>$\text{Corr}(Y_{ij}, Y_{i,j+t}) = \alpha^t$<br>for $t = 0, 1, 2, \ldots, n_i - j$ | $\hat{\alpha} = \frac{1}{(K_1 - p)\phi} \sum_{i=1}^{K} \sum_{j \leq n_i - 1} e_{ij} e_{i,j+1}$<br>$K_1 = \sum_{i=1}^{K} (n_i - 1)$ |

### *Dispersion Parameter*

The dispersion parameter $\phi$ is estimated by

$$\hat{\phi} = \frac{1}{N - p} \sum_{i=1}^{K} \sum_{j=1}^{n_i} e_{ij}^2$$

where $N = \sum_{i=1}^{K} n_i$ is the total number of measurements and $p$ is the number of regression parameters.

The square root of $\hat{\phi}$ is reported by PROC GENMOD as the scale parameter in the "Analysis of GEE Parameter Estimates Model-Based Standard Error Estimates" output table.

### Fitting Algorithm

The following is an algorithm for fitting the specified model using GEEs. Note that this is not in general a likelihood-based method of estimation, so that inferences based on likelihoods are not possible for GEE methods.

1. Compute an initial estimate of $\boldsymbol{\beta}$ with an ordinary generalized linear model assuming independence.

2. Compute the working correlations $\mathbf{R}$ based on the standardized residuals, the current $\boldsymbol{\beta}$, and the assumed structure of $\mathbf{R}$.

3. Compute an estimate of the covariance:

$$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \hat{\mathbf{R}}(\boldsymbol{\alpha}) \mathbf{A}_i^{\frac{1}{2}}$$

4. Update $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r + \left[ \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right]^{-1} \left[ \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right]$$

5. Iterate steps 2-4 until convergence

### Missing Data

Refer to Diggle, Liang, and Zeger (1994, Chapter 11) for a discussion of missing values in longitudinal data. Suppose that you intend to take measurements $Y_{i1}, \ldots, Y_{in}$ for the $i$th unit. Missing values for which $Y_{ij}$ are missing whenever $Y_{ik}$ is missing for all $j \geq k$ are called *dropouts*. Otherwise, missing values that occur intermixed with nonmissing values are *intermittent* missing values. The GENMOD procedure can estimate the working correlation from data containing both types of missing values using the *all available pairs* method, in which all nonmissing pairs of data are used in the moment estimators of the working correlation parameters defined previously.

For example, for the unstructured working correlation model,

$$\hat{\alpha}_{jk} = \frac{1}{(K' - p)\phi} \sum e_{ij} e_{ik}$$

where the sum is over the units that have nonmissing measurements at times $j$ and $k$, and $K'$ is the number of units with nonmissing measurements at $j$ and $k$. Estimates of the parameters for other working correlation types are computed in a similar manner, using available nonmissing pairs in the appropriate moment estimators.

The contribution of the $i$th unit to the parameter update equation is computed by omitting the elements of $(\mathbf{Y}_i - \boldsymbol{\mu}_i)$, the columns of $\mathbf{D}_i' = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}'$, and the rows and columns of $\mathbf{V}_i$ corresponding to missing measurements.

### Parameter Estimate Covariances

The *model-based* estimator of $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ is given by

$$\boldsymbol{\Sigma}_m(\hat{\boldsymbol{\beta}}) = \mathbf{I}_0^{-1}$$

where

$$\mathbf{I}_0 = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$$

This is the GEE equivalent of the inverse of the Fisher information matrix that is often used in generalized linear models as an estimator of the covariance estimate of the maximum likelihood estimator of $\boldsymbol{\beta}$. It is a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ if the mean model and the working correlation matrix are correctly specified.

The estimator

$$\boldsymbol{\Sigma}_e = \mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}$$

is called the *empirical*, or *robust*, estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ where

$$\mathbf{I}_1 = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_i^{-1} \mathrm{Cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$$

It has the property of being a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$, even if the working correlation matrix is misspecified, that is, if $\mathrm{Cov}(\mathbf{Y}_i) \neq \mathbf{V}_i$. In computing $\mathbf{M}$, $\boldsymbol{\beta}$ and $\phi$ are replaced by estimates, and $\mathrm{Cov}(\mathbf{Y}_i)$ is replaced by the estimate

$$(\mathbf{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}))(\mathbf{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}))'$$

### Multinomial GEEs

Lipsitz, Kim, and Zhao (1994) and Miller, Davis, and Landis (1993) describe how to extend GEEs to multinomial data. Currently, only the independent working correlation is available for multinomial models in PROC GENMOD.

### Alternating Logistic Regressions

If the responses are binary (that is, they take only two values), then there is an alternative method to account for the association among the measurements. The Alternating Logistic Regressions (ALR) algorithm of Carey, Zeger, and Diggle (1993) models the association between pairs of responses with log odds ratios, instead of with correlations, as ordinary GEEs do.

For binary data, the correlation between the *j*th and *k*th response is, by definition,

$$Corr(Y_{ij}, Y_{ik}) = \frac{\Pr(Y_{ij} = 1, Y_{ik} = 1) - \mu_{ij}\mu_{ik}}{\sqrt{\mu_{ij}(1-\mu_{ij})\mu_{ik}(1-\mu_{ik})}}$$

The joint probability in the numerator satisfies the following bounds, by elementary properties of probability, since $\mu_{ij} = \Pr(Y_{ij} = 1)$:

$$\max(0, \mu_{ij} + \mu_{ik} - 1) \leq \Pr(Y_{ij} = 1, Y_{ik} = 1) \leq \min(\mu_{ij}, \mu_{ik})$$

The correlation, therefore, is constrained to be within limits that depend in a complicated way on the means of the data.

The odds ratio, defined as

$$OR(Y_{ij}, Y_{ik}) = \frac{\Pr(Y_{ij} = 1, Y_{ik} = 1)\Pr(Y_{ij} = 0, Y_{ik} = 0)}{\Pr(Y_{ij} = 1, Y_{ik} = 0)\Pr(Y_{ij} = 0, Y_{ik} = 1)}$$

is not constrained by the means and is preferred, in some cases, to correlations for binary data.

The ALR algorithm seeks to model the logarithm of the odds ratio, $\gamma_{ijk} = \log(OR(Y_{ij}, Y_{ik}))$, as

$$\gamma_{ijk} = \mathbf{z}'_{ijk}\alpha$$

where $\alpha$ is a $q \times 1$ vector of regression parameters and $\mathbf{z}_{ijk}$ is a fixed, specified vector of coefficients.

The parameter $\gamma_{ijk}$ can take any value in $(-\infty, \infty)$ with $\gamma_{ijk} = 0$ corresponding to no association.

The log odds ratio, when modeled in this way with a regression model, can take different values in subgroups defined by $\mathbf{z}_{ijk}$. For example, $\mathbf{z}_{ijk}$ can define subgroups within clusters, or it can define 'block effects' between clusters.

You specify a GEE model for binary data using log odds ratios by specifying a model for the mean, as in ordinary GEEs, and a model for the log odds ratios. You can use any of the link functions appropriate for binary data in the model for the mean, such as logistic, probit, or complementary log-log. The ALR algorithm alternates between a GEE step to update the model for the mean and a logistic regression step to update the log odds ratio model. Upon convergence, the ALR algorithm provides estimates of the regression parameters for the mean, $\boldsymbol{\beta}$, the regression parameters for the log odds ratios, $\alpha$, their standard errors, and their covariances.

### Specifying Log Odds Ratio Models

Specifying a regression model for the log odds ratio requires you to specify rows of the *z*-matrix $\mathbf{z}_{ijk}$ for each cluster $i$ and each unique within-cluster pair $(j, k)$. The GENMOD procedure provides several methods of specifying $\mathbf{z}_{ijk}$. These are controlled by the LOGOR=*keyword* and associated options in the REPEATED statement. The supported keywords and the resulting log odds ratio models are described as follows.

EXCH — exchangeable log odds ratios. In this model, the log odds ratio is a constant for all clusters $i$ and pairs $(j,k)$. The parameter $\alpha$ is the common log odds ratio.

$$\mathbf{z}_{ijk} = 1 \quad \text{for all} \quad i,j,k$$

FULLCLUST — fully parameterized clusters. Each cluster is parameterized in the same way, and there is a parameter for each unique pair within clusters. If a complete cluster is of size $n$, then there are $\frac{n(n-1)}{2}$ parameters in the vector $\alpha$. For example, if a full cluster is of size 4, then there are $\frac{4\times 3}{2} = 6$ parameters, and the *z*-matrix is of the form

$$Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The elements of $\alpha$ correspond to log odds ratios for cluster pairs in the following order:

| Pair | Parameter |
|---|---|
| (1,2) | Alpha1 |
| (1,3) | Alpha2 |
| (1,4) | Alpha3 |
| (2.3) | Alpha4 |
| (2,4) | Alpha5 |
| (3,4) | Alpha6 |

LOGORVAR(*variable*) — log odds ratios by cluster. The argument *variable* is a variable name that defines the 'block effects' between clusters. The log odds ratios are constant within clusters, but they take a different value for each different value of the *variable*. For example, if Center is a variable in the input data set taking a different value for $k$ treatment centers, then specifying LOGOR=LOGORVAR(Center) requests a model with different log odds ratios for each of the $k$ centers, constant within center.

NESTK — *k*-nested log odds ratios. You must also specify the SUBCLUST=*variable* option to define subclusters within clusters. Within each cluster, PROC GENMOD computes a log odds ratio parameter for pairs having the same value of *variable* for both members of the pair and one log odds ratio parameter for each unique combination of different values of *variable*.

NEST1 — 1-nested log odds ratios. You must also specify the SUBCLUST=*variable* option to define subclusters within clusters. There are two log odds ratio parameters for this model. Pairs having the same value of *variable* correspond to one parameter; pairs having different values of *variable* correspond to the other parameter. For example, if clusters are hospitals and subclusters are wards within hospitals, then patients within the same ward have one log odds ratio parameter, and patients from different wards have the other parameter.

ZFULL — specifies the full *z*-matrix. You must also specify a SAS data set containing the *z*-matrix with the ZDATA=*data-set-name* option. Each observation in the data set corresponds to one row of the *z*-matrix. You must specify the ZDATA data set as if all clusters are complete, that is, as if all clusters are the same size and there are no missing observations. The ZDATA data set has $K[n_{max}(n_{max}-1)/2]$ observations, where $K$ is the number of clusters and $n_{max}$ is the maximum cluster size. If the members of cluster $i$ are ordered as $1, 2, \cdots, n$, then the rows of the *z*-matrix must be specified for pairs in the order $(1,2), (1,3), \cdots, (1,n), (2,3), \cdots, (2,n), \cdots, (n-1,n)$. The variables specified in the REPEATED statement for the SUBJECT effect must also be present in the ZDATA= data set to identify clusters. You must specify variables in the data set that define the columns of the *z*-matrix by the ZROW=*variable-list* option. If there are $q$ columns, ($q$ variables in *variable-list*), then there are $q$ log odds ratio parameters. You can optionally specify variables indicating the cluster pairs corresponding to each row of the *z*-matrix with the YPAIR=(*variable1, variable2*) option. If you specify this option, the data from the ZDATA data set is sorted within each cluster by *variable1* and *variable2*. See Example 29.6 for an example of specifying a full *z*-matrix.

ZREP — replicated *z*-matrix. You specify *z*-matrix data exactly as you do for the ZFULL case, except that you specify only one complete cluster. The *z*-matrix for the one cluster is replicated for each cluster. The number of observations in the ZDATA data set is $\frac{n_{max}(n_{max}-1)}{2}$, where $n_{max}$ is the size of a complete cluster (a cluster with no missing observations). See Example 29.6 for an example of specifying a replicated *z*-matrix.

ZREP(matrix) — direct input of the replicated *z*-matrix. You specify the *z*-matrix for one cluster with the syntax LOGOR=ZREP ( $(y_1 \;\; y_2) z_1 \;\; z_2 \cdots z_q, \cdots$ ), where $y_1$ and $y_2$ are numbers representing a pair of observations and the values

$z_1, z_2, \cdots, z_q$ make up the corresponding row of the *z*-matrix. The number of rows specified is $\frac{n_{max}(n_{max}-1)}{2}$, where $n_{max}$ is the size of a complete cluster (a cluster with no missing observations). For example,

```
LOGOR =  ZREP((1 2) 1 0,
              (1 3) 1 0,
              (1 4) 1 0,
              (2 3) 1 1,
              (2 4) 1 1,
              (3 4) 1 1)
```

specifies the $\frac{4\times 3}{2} = 6$ rows of the *z*-matrix for a cluster of size 4 with $q = 2$ log odds ratio parameters. The log odds ratio for pairs (1 2), (1 3), (1 4) is $\alpha_1$, and the log odds ratio for pairs (2 3), (2 4), (3 4) is $\alpha_1 + \alpha_2$.

### Generalized Score Statistics

Boos (1992) and Rotnitzky and Jewell (1990) describe score tests applicable to testing $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ in GEEs, where $\mathbf{L}$ is a user-specified $r \times p$ contrast matrix or a contrast for a Type 3 test of hypothesis.

Let $\tilde{\boldsymbol{\beta}}$ be the regression parameters resulting from solving the GEE under the restricted model $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, and let $\mathbf{S}(\tilde{\boldsymbol{\beta}})$ be the generalized estimating equation values at $\tilde{\boldsymbol{\beta}}$.

The generalized score statistic is

$$T = \mathbf{S}(\tilde{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_m\mathbf{L}'(\mathbf{L}\boldsymbol{\Sigma}_e\mathbf{L}')^{-1}\mathbf{L}\boldsymbol{\Sigma}_m\mathbf{S}(\tilde{\boldsymbol{\beta}})$$

where $\boldsymbol{\Sigma}_m$ is the model-based covariance estimate and $\boldsymbol{\Sigma}_e$ is the empirical covariance estimate. The *p*-values for $T$ are computed based on the chi-square distribution with *r* degrees of freedom.

## Displayed Output

The following output is produced by the GENMOD procedure. Note that some of the tables are optional and appear only in conjunction with the REPEATED statement and its options or with options in the MODEL statement. For details, see the section "ODS Table Names" on page 1437.

### Model Information

PROC GENMOD displays the following model information:

- data set name
- response distribution
- link function
- response variable name
- offset variable name

- frequency variable name
- scale weight variable name
- number of observations used
- number of events if events/trials format is used for response
- number of trials if events/trials format is used for response
- sum of frequency weights
- number of missing values in data set
- number of invalid observations (for example, negative or 0 response values with gamma distribution or number of observations with events greater than trials with binomial distribution)

### *Class Level Information*

If you use classification variables in the model, PROC GENMOD displays the levels of classification variables specified in the CLASS statement and in the MODEL statement. The levels are displayed in the same sorted order used to generate columns in the design matrix.

### *Response Profile*

If you specify an ordinal model for the multinomial distribution, a table titled "Response Profile" is displayed containing the ordered values of the response variable and the number of occurrences of the values used in the model.

### *Iteration History for Parameter Estimates*

If you specify the ITPRINT model option, PROC GENMOD displays a table containing the following for each iteration in the Newton-Raphson procedure for model fitting:

- iteration number
- ridge value
- log likelihood
- values of all parameters in the model

### *Criteria for Assessing Goodness of Fit*

PROC GENMOD displays the following criteria for assessing goodness of fit:

- degrees of freedom for deviance and Pearson's chi-square, equal to the number of observations minus the number of regression parameters estimated
- deviance
- deviance divided by degrees of freedom
- scaled deviance
- scaled deviance divided by degrees of freedom
- Pearson's chi-square
- Pearson's chi-square divided by degrees of freedom

- scaled Pearson's chi-square
- scaled Pearson's chi-square divided by degrees of freedom
- log likelihood

### Last Evaluation of the Gradient

If you specify the model option ITPRINT, the GENMOD procedure displays the last evaluation of the gradient vector.

### Last Evaluation of the Hessian

If you specify the model option ITPRINT, the GENMOD procedure displays the last evaluation of the Hessian matrix.

### Analysis of (Initial) Parameter Estimates

The "Analysis of (Initial) Parameter Estimates" table contains the results from fitting a generalized linear model to the data. If you specify the REPEATED statement, these GLM parameter estimates are used as initial values for the GEE solution. For each parameter in the model, PROC GENMOD displays the following:

- the parameter name
  - the variable name for continuous regression variables
  - the variable name and level for classification variables and interactions involving classification variables
  - SCALE for the scale variable related to the dispersion parameter
- degrees of freedom for the parameter
- estimate value
- standard error
- Wald chi-square value
- $p$-value based on the chi-square distribution
- confidence limits (Wald or profile likelihood) for parameters

### Estimated Covariance Matrix

If you specify the model option COVB, the GENMOD procedure displays the estimated covariance matrix, defined as the inverse of the information matrix at the final iteration. This is based on the expected information matrix if the EXPECTED option is specified in the MODEL statement. Otherwise, it is based on the Hessian matrix used at the final iteration. This is, by default, the observed Hessian unless altered by the SCORING option in the MODEL statement.

### Estimated Correlation Matrix

If you specify the CORRB model option, PROC GENMOD displays the estimated correlation matrix. This is based on the expected information matrix if the EXPECTED option is specified in the MODEL statement. Otherwise, it is based on the Hessian matrix used at the final iteration. This is, by default, the observed Hessian unless altered by the SCORING option in the MODEL statement.

### *Iteration History for LR Confidence Intervals*

If you specify the ITPRINT and LRCI model options, PROC GENMOD displays an iteration history table for profile likelihood-based confidence intervals. For each parameter in the model, PROC GENMOD displays the following:

- parameter identification number
- iteration number
- log likelihood value
- parameter values

### *Likelihood Ratio-Based Confidence Intervals for Parameters*

If you specify the LRCI and the ITPRINT options in the MODEL statement, a table is displayed summarizing profile likelihood-based confidence intervals for all parameters. The table contains the following for each parameter in the model:

- confidence coefficient
- parameter identification number
- lower and upper endpoints of confidence intervals for the parameter
- values of all other parameters at the solution

### *LR Statistics for Type 1 Analysis*

If you specify the TYPE1 model option, a table containing the following is displayed for each effect in the model:

- name of effect
- deviance for the model including the effect and all previous effects
- degrees of freedom for the effect
- likelihood ratio statistic for testing the significance of the effect
- $p$-value computed from the chi-square distribution with effect degrees of freedom

If you specify either the SCALE=DEVIANCE or SCALE=PEARSON option in the MODEL statement, columns containing the following are displayed:

- name of effect
- deviance for the model including the effect and all previous effects
- numerator degrees of freedom
- denominator degrees of freedom
- chi-square statistic for testing the significance of the effect
- $p$-value computed from the chi-square distribution with numerator degrees of freedom
- $F$ statistic for testing the significance of the effect
- $p$-value based on the $F$ distribution

### Iteration History for Type 3 Contrasts

If you specify the model options ITPRINT and TYPE3, an iteration history table is displayed for fitting the model with Type 3 contrast constraints for each effect. The table contains the following:

- effect name
- iteration number
- ridge value
- log likelihood
- values of all parameters

### LR Statistics for Type 3 Analysis

If you specify the TYPE3 model option, a table containing the following is displayed for each effect in the model:

- name of the effect
- likelihood ratio statistic for testing the significance of the effect
- degrees of freedom for effect
- $p$-value computed from the chi-square distribution

If you specify either the SCALE=DEVIANCE or SCALE=PEARSON option in the MODEL statement, columns containing the following are displayed:

- name of the effect
- likelihood ratio statistic for testing the significance of the effect
- $F$ statistic for testing the significance of the effect
- numerator degrees of freedom
- denominator degrees of freedom
- $p$-value based on the $F$ distribution
- $p$-value computed from the chi-square distribution with numerator degrees of freedom

### Wald Statistics for Type 3 Analysis

If you specify the TYPE3 and WALD model options, a table containing the following is displayed for each effect in the model:

- name of effect
- degrees of freedom for effect
- Wald statistic for testing the significance of the effect
- $p$-value computed from the chi-square distribution

### *Parameter Information*

If you specify the ITPRINT, COVB, CORRB, WALDCI, or LRCI option in the MODEL statement, or if you specify a CONTRAST statement, a table is displayed that identifies parameters with numbers, rather than names, for use in tables and matrices where a compact identifier for parameters is helpful. For each parameter, the table contains the following:

- a number that identifies the parameter
- the parameter name, including level information for effects containing classification variables

### *Observation Statistics*

If you specify the OBSTATS option in the MODEL statement, PROC GENMOD displays a table containing miscellaneous statistics. For each observation in the input data set, the following are displayed:

- the value of the response variable, denoted by the variable name
- the predicted value of the mean, denoted by PRED
- the value of the linear predictor, denoted by XBETA. The value of an OFFSET variable is not added to the linear predictor.
- the estimated standard error of the linear predictor, denoted by STD
- the value of the negative of the weight in the Hessian matrix at the final iteration, denoted by HESSWGT. This is the expected weight if the EXPECTED option is specified in the MODEL statement. Otherwise, it is the weight used in the final iteration. That is, it is the observed weight unless the SCORING= option has been specified.
- approximate lower and upper endpoints for a confidence interval for the predicted value of the mean, denoted by LOWER and UPPER
- raw residual, denoted by RESRAW
- Pearson residual, denoted by RESCHI
- deviance residual, denoted by RESDEV
- standardized Pearson residual, denoted by STDRESCHI
- standardized deviance residual, denoted by STDRESDEV
- likelihood residual, denoted by RESLIK

### *ESTIMATE Statement Results*

If you specify a REPEATED statement, the ESTIMATE statement results apply to the specified GEE model. Otherwise, they apply to the specified generalized linear model.

The following are displayed for each ESTIMATE statement:

- contrast label
- estimated value of the contrast
- standard error of the estimate
- significance level $\alpha$
- $(1 - \alpha) \times 100\%$ confidence intervals for contrast
- Wald chi-square statistic for the contrast
- $p$-value computed from the chi-square distribution

If you specify the EXP option, an additional row is displayed with statistics for the exponentiated value of the contrast.

### CONTRAST Coefficients

If you specify the CONTRAST or ESTIMATE statement and you specify the E option, a table titled "Coefficients For Contrast *label*" is displayed, where *label* is the label specified in the CONTRAST statement. The table contains the following:

- the contrast label
- the rows of the contrast matrix

### Iteration History for Contrasts

If you specify the ITPRINT option, an iteration history table is displayed for fitting the model with contrast constraints for each effect. The table contains the following for each contrast defined in a CONTRAST statement:

- contrast label
- iteration number
- ridge value
- log likelihood
- values of all parameters

### CONTRAST Statement Results

If you specify a REPEATED statement, the CONTRAST statement results apply to the specified GEE model. Otherwise, they apply to the specified generalized linear model.

The following are displayed for each CONTRAST statement:

- contrast label
- degrees of freedom for the contrast
- likelihood ratio, score, or Wald statistic for testing the significance of the contrast. Score statistics are used in GEE models, likelihood ratio statistics are used in generalized linear models, and Wald statistics are used in both.
- $p$-value computed from the chi-square distribution
- the type of statistic computed for this contrast: Wald, LR, or score

If you specify either the SCALE=DEVIANCE or SCALE=PEARSON option for generalized linear models, columns containing the following are displayed:

- contrast label
- likelihood ratio statistic for testing the significance of the contrast
- $F$ statistic for testing the significance of the contrast
- numerator degrees of freedom
- denominator degrees of freedom
- $p$-value based on the $F$ distribution
- $p$-value computed from the chi-square distribution with numerator degrees of freedom

### *LSMEANS Coefficients*

If you specify the LSMEANS statement and you specify the E option, a table titled "Coefficients for *effect* Least Squares Means" is displayed, where *effect* is the effect specified in the LSMEANS statement. The table contains the following:

- the effect names
- the rows of least squares means coefficients

### *Least Squares Means*

If you specify the LSMEANS statement a table titled "Least Squares Means" is displayed. The table contains the following:

- the effect names
- for each level of each effect,
  - the least squares mean estimate
  - standard error
  - chi-square value
  - $p$-value computed from the chi-square distribution

If you specify the DIFF option, a table titled "Differences of Least Squares Means" is displayed containing corresponding statistics for the differences between the least squares means for the levels of each effect.

### GEE Model Information

If you specify the REPEATED statement, the following are displayed:

- correlation structure of the working correlation matrix or the log odds ratio structure
- within-subject effect
- subject effect
- number of clusters
- correlation matrix dimension
- minimum and maximum cluster size

### Log Odds Ratio Parameter Information

If you specify the REPEATED statement and specify a log odds ratio model for binary data with the LOGOR= option, then a table is displayed showing the correspondence between data pairs and log odds ratio model parameters.

### Iteration History for GEE Parameter Estimates

If you specify the REPEATED statement and the MODEL statement option ITPRINT, an iteration history table for GEE parameter estimates is displayed. The table contains the following:

- parameter identification number
- iteration number
- values of all parameters

### Last Evaluation of the Generalized Gradient and Hessian

If you specify the REPEATED statement and select ITPRINT as a model option, PROC GENMOD displays the last generalized gradient and Hessian matrix in the GEE iterative parameter estimation process.

### GEE Parameter Estimate Covariance Matrices

If you specify the REPEATED statement and the COVB option, PROC GENMOD displays both model-based and empirical parameter estimate covariance matrices.

### GEE Parameter Estimate Correlation Matrices

If you specify the REPEATED statement and the CORRB option, PROC GENMOD displays both model-based and empirical parameter estimate covariance matrices.

### GEE Working Correlation Matrix

If you specify the REPEATED statement and the CORRW option, PROC GENMOD displays the exchangeable working correlation matrix.

### Analysis of GEE Parameter Estimates

If you specify the REPEATED statement, PROC GENMOD uses empirical standard error estimates to compute and display the following for each parameter in the model:

- the parameter name
  - the variable name for continuous regression variables
  - the variable name and level for classification variables and interactions involving classification variables
  - "Scale" for the scale variable related to the dispersion parameter
- parameter estimate
- standard error
- 95% confidence interval
- $Z$ score and $p$-value

If you specify the MODELSE option in the REPEATED statement, a table based on model-based standard errors is also produced.

### GEE Observation Statistics

If you specify the OBSTATS option in the REPEATED statement, PROC GENMOD displays a table containing miscellaneous statistics. For each observation in the input data set, the following are displayed:

- the value of the response variable and all other variables in the model, denoted by the variable names
- the predicted value of the mean, denoted by PRED
- the value of the linear predictor, denoted by XBETA
- the standard error of the linear predictor, denoted by STD
- confidence limits for the predicted values, denoted by LOWER and UPPER
- raw residual, denoted by RESRAW
- Pearson residual, denoted by RESCHI

## ODS Table Names

PROC GENMOD assigns a name to each table that it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

**Table 29.3.** ODS Tables Produced in PROC GENMOD

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| ClassLevels | Class variable levels | CLASS | default |
| Contrasts | Tests of contrasts | CONTRAST | default |
| ContrastCoef | Contrast coefficients | CONTRAST | E |
| ConvergenceStatus | Convergence status | MODEL | default |
| CorrB | Parameter estimate correlation matrix | MODEL | CORRB |
| CovB | Parameter estimate covariance matrix | MODEL | COVB |
| Estimates | Estimates of contrasts | ESTIMATE | default |
| EstimateCoef | Contrast coefficients | ESTIMATE | E |
| GEEEmpPEst | GEE parameter estimates with empirical standard errors | REPEATED | default |
| GEELogORInfo | GEE log odds ratio model information | REPEATED | LOGOR= |
| GEEModInfo | GEE model information | REPEATED | default |
| GEEModPEst | GEE parameter estimates with model-based standard errors | REPEATED | MODELSE |
| GEENCorr | GEE model-based correlation matrix | REPEATED | MCORRB |
| GEENCov | GEE model-based covariance matrix | REPEATED | MCOVB |
| GEERCorr | GEE empirical correlation matrix | REPEATED | ECORRB |
| GEERCov | GEE empirical covariance matrix | REPEATED | ECOVB |
| GEEWCorr | GEE working correlation matrix | REPEATED | CORRW |
| IterContrasts | Iteration history for contrasts | MODEL CONTRAST | ITPRINT |
| IterLRCI | Iteration history for likelihood ratio confidence intervals | MODEL | LRCI ITPRINT |
| IterParms | Iteration history for parameter estimates | MODEL | ITPRINT |
| IterParmsGEE | Iteration history for GEE parameter estimates | MODEL REPEATED | ITPRINT |
| IterType3 | Iteration history for Type 3 statistics | MODEL | TYPE3 ITPRINT |
| LRCI | Likelihood ratio confidence intervals | MODEL | LRCI ITPRINT |
| LSMeanCoef | Coefficients for least squares means | LSMEANS | E |

**Table 29.3.** (continued)

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| LSMeanDiffs | Least squares means differences | LSMEANS | DIFF |
| LSMeans | Least squares means | LSMEANS | default |
| LagrangeStatistics | Lagrange statistics | MODEL | NOINT \| NOSCALE |
| LastGEEGrad | Last evaluation of the generalized gradient and Hessian | MODEL REPEATED | ITPRINT |
| LastGradHess | Last evaluation of the gradient and Hessian | MODEL | ITPRINT |
| LinDep | Linearly dependent rows of contrasts | CONTRAST * | default |
| ModelInfo | Model information | MODEL | default |
| Modelfit | Goodness-of-fit statistics | MODEL | default |
| NonEst | Nonestimable rows of contrasts | CONTRAST * | default |
| ObStats | Observation-wise statistics | MODEL | OBSTATS \| CL \| PREDICTED \| RESIDUALS \| XVARS |
| ParameterEstimates | Parameter estimates | MODEL | default |
| ParmInfo | Parameter indices | MODEL * | default |
| ResponseProfiles | Frequency counts for multinomial models | MODEL | DIST=MULTINOMIAL |
| Type1 | Type 1 tests | MODEL | TYPE1 |
| Type3 | Type 3 tests | MODEL | TYPE3 |

*Depends on data.

# Examples

The following examples illustrate some of the capabilities of the GENMOD procedure. These are not intended to represent definitive analyses of the data sets presented here. You should refer to the texts cited in the "References" section on page 1462 for guidance on complete analysis of data using generalized linear models.

## Example 29.1. Logistic Regression

In an experiment comparing the effects of five different drugs, each drug is tested on a number of different subjects. The outcome of each experiment is the presence or absence of a positive response in a subject. The following artificial data represent the number of responses r in the n subjects for the five different drugs, labeled A through E. The response is measured for different levels of a continuous covariate x for each drug. The drug type and the continuous covariate x are explanatory variables in this experiment. The number of responses r is modeled as a binomial random variable for each combination of the explanatory variable values, with the binomial number of trials parameter equal to the number of subjects n and the binomial probability equal to the probability of a response. The following DATA step creates the data set.

```
data drug;
   input drug$ x r n @@;
   datalines;
A  .1   1  10   A  .23  2  12   A  .67  1   9
B  .2   3  13   B  .3   4  15   B  .45  5  16   B  .78  5  13
C  .04  0  10   C  .15  0  11   C  .56  1  12   C  .7   2  12
D  .34  5  10   D  .6   5   9   D  .7   8  10
E  .2  12  20   E  .34 15  20   E  .56 13  15   E  .8  17  20
;
```

A logistic regression for these data is a generalized linear model with response equal to the binomial proportion r/n. The probability distribution is binomial, and the link function is logit. For these data, drug and x are explanatory variables. The probit and the complementary log-log link functions are also appropriate for binomial data.

PROC GENMOD performs a logistic regression on the data in the following SAS statements:

```
proc genmod data=drug;
   class drug;
   model r/n = x drug / dist = bin
                        link = logit
                        lrci
                        ;
run;
```

Since these data are binomial, you use the events/trials syntax to specify the response in the MODEL statement. Profile likelihood confidence intervals for the regression parameters are computed using the LRCI option.

General model and data information is produced in Output 29.1.1.

**Output 29.1.1.** Model Information

The GENMOD Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.DRUG |
| Distribution | Binomial |
| Link Function | Logit |
| Response Variable (Events) | r |
| Response Variable (Trials) | n |
| Observations Used | 18 |
| Number Of Events | 99 |
| Number Of Trials | 237 |

The five levels of the CLASS variable DRUG are displayed in Output 29.1.2.

**Output 29.1.2.** Class Variable Levels

The GENMOD Procedure

Class Level Information

| Class | Levels | Values |
|---|---|---|
| drug | 5 | A B C D E |

In the "Criteria For Assessing Goodness Of Fit" table displayed in Output 29.1.3, the value of the deviance divided by its degrees of freedom is less than 1. A $p$-value is not computed for the deviance; however, a deviance that is approximately equal to its degrees of freedom is a possible indication of a good model fit. Asymptotic distribution theory applies to binomial data as the number of binomial trials parameter n becomes large for each combination of explanatory variables. McCullagh and Nelder (1989) caution against the use of the deviance alone to assess model fit. The model fit for each observation should be assessed by examination of residuals. The OBSTATS option in the MODEL statement produces a table of residuals and other useful statistics for each observation.

**Output 29.1.3.** Goodness of Fit Criteria

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 12 | 5.2751 | 0.4396 |
| Scaled Deviance | 12 | 5.2751 | 0.4396 |
| Pearson Chi-Square | 12 | 4.5133 | 0.3761 |
| Scaled Pearson X2 | 12 | 4.5133 | 0.3761 |
| Log Likelihood | | -114.7732 | |

In the "Analysis Of Parameter Estimates" table displayed in Output 29.1.4, chi-square values for the explanatory variables indicate that the parameter values other than the intercept term are all significant. The scale parameter is set to 1 for the binomial distribution. When you perform an overdispersion analysis, the value of the overdispersion parameter is indicated here. See the the section "Overdispersion" on page 1410 for a discussion of overdispersion.

**Output 29.1.4.** Parameter Estimates

The GENMOD Procedure

Analysis Of Parameter Estimates

| Parameter | | DF | Estimate | Standard Error | Likelihood Ratio 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|
| Intercept | | 1 | 0.2792 | 0.4196 | -0.5336 | 1.1190 | 0.44 | 0.5057 |
| x | | 1 | 1.9794 | 0.7660 | 0.5038 | 3.5206 | 6.68 | 0.0098 |
| drug | A | 1 | -2.8955 | 0.6092 | -4.2280 | -1.7909 | 22.59 | <.0001 |
| drug | B | 1 | -2.0162 | 0.4052 | -2.8375 | -1.2435 | 24.76 | <.0001 |
| drug | C | 1 | -3.7952 | 0.6655 | -5.3111 | -2.6261 | 32.53 | <.0001 |
| drug | D | 1 | -0.8548 | 0.4838 | -1.8072 | 0.1028 | 3.12 | 0.0773 |
| drug | E | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| Scale | | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

NOTE: The scale parameter was held fixed.

The preceding table contains the profile likelihood confidence intervals for the explanatory variable parameters requested with the LRCI option. Wald confidence intervals are displayed by default. Profile likelihood confidence intervals are considered to be more accurate than Wald intervals (refer to Aitkin et al. 1989), especially with small sample sizes. You can specify the confidence coefficient with the ALPHA= option in the MODEL statement. The default value of 0.05, corresponding to 95% confidence limits, is used here. See the section "Confidence Intervals for Parameters" on page 1415 for a discussion of profile likelihood confidence intervals.

## Example 29.2. Normal Regression, Log Link

Consider the following data, where x is an explanatory variable, and y is the response variable. It appears that y varies nonlinearly with x and that the variance is approximately constant. A normal distribution with a log link function is chosen to model these data; that is, $\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$ so that $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$.

```
data nor;
   input x y;
   datalines;
0 5
0 7
0 9
1 7
1 10
1 8
2 11
2 9
3 16
3 13
3 14
4 25
4 24
5 34
5 32
5 30
;
```

The following SAS statements produce the analysis with the normal distribution and log link:

```
proc genmod data=nor;
   model y = x / dist = normal
                 link = log
                 ;
   output out = Residuals
          pred = Pred
          resraw = Resraw
          reschi = Reschi
          resdev = Resdev
          stdreschi = Stdreschi
          stdresdev = Stdresdev
          reslik = Reslik;
   proc print data=Residuals;
run;
```

The OUTPUT statement is specified to produce a data set that contains predicted values and residuals for each observation. This data set can be useful for further analysis, such as residual plotting.

The output from these statements is displayed in Output 29.2.1.

**Output 29.2.1.** Log Linked Normal Regression

The GENMOD Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.NOR |
| Distribution | Normal |
| Link Function | Log |
| Dependent Variable | y |
| Observations Used | 16 |

Criteria For Assessing Goodness Of Fit

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 14 | 52.3000 | 3.7357 |
| Scaled Deviance | 14 | 16.0000 | 1.1429 |
| Pearson Chi-Square | 14 | 52.3000 | 3.7357 |
| Scaled Pearson X2 | 14 | 16.0000 | 1.1429 |
| Log Likelihood | | -32.1783 | |

Analysis Of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 1.7214 | 0.0894 | 1.5461 | 1.8966 | 370.76 | <.0001 |
| x | 1 | 0.3496 | 0.0206 | 0.3091 | 0.3901 | 286.64 | <.0001 |
| Scale | 1 | 1.8080 | 0.3196 | 1.2786 | 2.5566 | | |

NOTE: The scale parameter was estimated by maximum likelihood.

The PROC GENMOD scale parameter, in the case of the normal distribution, is the standard deviation. By default, the scale parameter is estimated by maximum likelihood. You can specify a fixed standard deviation by using the NOSCALE and SCALE= options in the MODEL statement.

**Output 29.2.2.** Data Set of Predicted Values and Residuals

| Obs | x | y | Pred | Reschi | Resdev | Resraw | Stdreschi | Stdresdev | Reslik |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 5 | 5.5921 | -0.59212 | -0.59212 | -0.59212 | -0.34036 | -0.34036 | -0.34036 |
| 2 | 0 | 7 | 5.5921 | 1.40788 | 1.40788 | 1.40788 | 0.80928 | 0.80928 | 0.80928 |
| 3 | 0 | 9 | 5.5921 | 3.40788 | 3.40788 | 3.40788 | 1.95892 | 1.95892 | 1.95892 |
| 4 | 1 | 7 | 7.9324 | -0.93243 | -0.93243 | -0.93243 | -0.54093 | -0.54093 | -0.54093 |
| 5 | 1 | 10 | 7.9324 | 2.06757 | 2.06757 | 2.06757 | 1.19947 | 1.19947 | 1.19947 |
| 6 | 1 | 8 | 7.9324 | 0.06757 | 0.06757 | 0.06757 | 0.03920 | 0.03920 | 0.03920 |
| 7 | 2 | 11 | 11.2522 | -0.25217 | -0.25217 | -0.25217 | -0.14686 | -0.14686 | -0.14686 |
| 8 | 2 | 9 | 11.2522 | -2.25217 | -2.25217 | -2.25217 | -1.31166 | -1.31166 | -1.31166 |
| 9 | 3 | 16 | 15.9612 | 0.03878 | 0.03878 | 0.03878 | 0.02249 | 0.02249 | 0.02249 |
| 10 | 3 | 13 | 15.9612 | -2.96122 | -2.96122 | -2.96122 | -1.71738 | -1.71738 | -1.71738 |
| 11 | 3 | 14 | 15.9612 | -1.96122 | -1.96122 | -1.96122 | -1.13743 | -1.13743 | -1.13743 |
| 12 | 4 | 25 | 22.6410 | 2.35897 | 2.35897 | 2.35897 | 1.37252 | 1.37252 | 1.37252 |
| 13 | 4 | 24 | 22.6410 | 1.35897 | 1.35897 | 1.35897 | 0.79069 | 0.79069 | 0.79069 |
| 14 | 5 | 34 | 32.1163 | 1.88366 | 1.88366 | 1.88366 | 1.22914 | 1.22914 | 1.22914 |
| 15 | 5 | 32 | 32.1163 | -0.11634 | -0.11634 | -0.11634 | -0.07592 | -0.07592 | -0.07592 |
| 16 | 5 | 30 | 32.1163 | -2.11634 | -2.11634 | -2.11634 | -1.38098 | -1.38098 | -1.38098 |

The data set of predicted values and residuals (Output 29.2.2) is created by the OUTPUT statement. With this data set, you can construct residual plots using the GPLOT

procedure to aid in assessing model fit. Note that raw, Pearson, and deviance residuals are equal in this example. This is a characteristic of the normal distribution and is not true in general for other distributions.

## Example 29.3. Gamma Distribution Applied to Life Data

Life data are sometimes modeled with the gamma distribution. Although PROC GENMOD does not analyze censored data or provide other useful lifetime distributions such as the Weibull or lognormal, it can be used for modeling complete (uncensored) data with the gamma distribution, and it can provide a statistical test for the exponential distribution against other gamma distribution alternatives. Refer to Lawless (1982) or Nelson (1982) for applications of the gamma distribution to life data.

The following data represent failure times of machine parts, some of which are manufactured by manufacturer A and some by manufacturer B.

```
data A;
   input lifetime@@ ;
   mfg = 'A';
   datalines;
620  470  260  89   388  242
103  100  39   460  284  1285
218  393  106  158  152  477
403  103  69   158  818  947
399  1274 32   12   134  660
548  381  203  871  193  531
317  85   1410 250  41   1101
32   421  32   343  376  1512
1792 47   95   76   515  72
1585 253  6    860  89   1055
537  101  385  176  11   565
164  16   1267 352  160  195
1279 356  751  500  803  560
151  24   689  1119 1733 2194
763  555  14   45   776  1
;

data B;
   input lifetime@@ ;
   mfg = 'B';
   datalines;
1747 945  12   1453 14   150
20   41   35   69   195  89
1090 1868 294  96   618  44
142  892  1307 310  230  30
403  860  23   406  1054 1935
561  348  130  13   230  250
317  304  79   1793 536  12
9    256  201  733  510  660
122  27   273  1231 182  289
667  761  1096 43   44   87
```

```
405  998  1409 61   278  407
113  25   940  28   848  41
646  575  219  303  304  38
195  1061 174  377  388  10
246  323  198  234  39   308
55   729  813  1216 1618 539
6    1566 459  946  764  794
35   181  147  116  141  19
380  609  546
;

data lifdat;
   set A B;
run;
```

The following SAS statements use PROC GENMOD to compute Type 3 statistics to test for differences between the two manufacturers in machine part life. Type 3 statistics are identical to Type 1 statistics in this case, since there is only one effect in the model. The log link function is selected to ensure that the mean is positive.

```
proc genmod data = lifdat;
   class mfg;
   model lifetime = mfg / dist=gamma
                          link=log
                          type3;
run;
```

The output from these statements is displayed in Output 29.3.1.

**Output 29.3.1.** Gamma Model of Life Data

**The GENMOD Procedure**

**Model Information**

| | |
|---|---|
| Data Set | WORK.LIFDAT |
| Distribution | Gamma |
| Link Function | Log |
| Dependent Variable | lifetime |
| Observations Used | 201 |

**Class Level Information**

| Class | Levels | Values |
|---|---|---|
| mfg | 2 | A B |

**Criteria For Assessing Goodness Of Fit**

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 199 | 287.0591 | 1.4425 |
| Scaled Deviance | 199 | 237.5335 | 1.1936 |
| Pearson Chi-Square | 199 | 211.6870 | 1.0638 |
| Scaled Pearson X2 | 199 | 175.1652 | 0.8802 |
| Log Likelihood | | -1432.4177 | |

**Analysis Of Parameter Estimates**

| Parameter | | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|
| Intercept | | 1 | 6.1302 | 0.1043 | 5.9257 | 6.3347 | 3451.61 | <.0001 |
| mfg | A | 1 | 0.0199 | 0.1559 | -0.2857 | 0.3255 | 0.02 | 0.8985 |
| mfg | B | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| Scale | | 1 | 0.8275 | 0.0714 | 0.6987 | 0.9800 | | |

NOTE: The scale parameter was estimated by maximum likelihood.

**LR Statistics For Type 3 Analysis**

| Source | DF | Chi-Square | Pr > ChiSq |
|---|---|---|---|
| mfg | 1 | 0.02 | 0.8985 |

The $p$-value of 0.8985 for the chi-square statistic in the Type 3 table indicates that there is no significant difference in the part life for the two manufacturers.

Using the following statements, you can refit the model without using the manufacturer as an effect. The LRCI option in the MODEL statement is specified to compute profile likelihood confidence intervals for the mean life and scale parameters.

```
proc genmod data = lifdat;
   model lifetime = / dist=gamma
                      link=log
                      lrci;
run;
```

**Output 29.3.2.** Refitting of the Gamma Model: Omitting the mfg Effect

```
                          The GENMOD Procedure

                     Analysis Of Parameter Estimates

                                        Likelihood Ratio
                                 Standard    95% Confidence       Chi-
Parameter  DF  Estimate     Error        Limits          Square  Pr > ChiSq

Intercept   1    6.1391    0.0775    5.9904    6.2956  6268.10      <.0001
Scale       1    0.8274    0.0714    0.6959    0.9762

NOTE: The scale parameter was estimated by maximum likelihood.
```

The intercept is the estimated log mean of the fitted gamma distribution, so that the mean life of the parts is

$$\mu = \exp(\text{INTERCEPT}) = \exp(6.1391) = 463.64$$

The SCALE parameter used in PROC GENMOD is the inverse of the gamma dispersion parameter, and it is sometimes called the gamma *index parameter*. See the "Response Probability Distributions" section on page 1402 for the definition of the gamma probability density function. A value of 1 for the index parameter corresponds to the exponential distribution . The estimated value of the scale parameter is 0.8274. The 95% profile likelihood confidence interval for the scale parameter is (0.6959, 0.9762), which does not contain 1. The hypothesis of an exponential distribution for the data is, therefore, rejected at the 0.05 level. A confidence interval for the mean life is

$$(\exp(5.99), \exp(6.30)) = (399.57, 542.18)$$

## Example 29.4. Ordinal Model for Multinomial Data

This example illustrates how you can use the GENMOD procedure to fit a model to data measured on an ordinal scale. The following statements create a SAS data set called icecream. The data set contains the results of a hypothetical taste test of three brands of ice cream. The three brands are rated for taste on a five point scale from very good (vg) to very bad (vb). An analysis is performed to assess the differences in the ratings for the three brands. The variable taste contains the ratings and brand contains the brands tested. The variable count contains the number of testers rating each brand in each category.

The following statements create the icecream data set.

```
data icecream;
   input count brand$ taste$;
   datalines;
70  ice1 vg
71  ice1 g
151 ice1 m
30  ice1 b
46  ice1 vb
20  ice2 vg
36  ice2 g
130 ice2 m
74  ice2 b
70  ice2 vb
50  ice3 vg
55  ice3 g
140 ice3 m
52  ice3 b
50  ice3 vb
;
```

The following statements fit a cumulative logit model to the ordinal data with the variable taste as the response and the variable brand as a covariate. The variable count is used as a FREQ variable.

```
proc genmod  rorder=data;
   freq count;
   class brand;
   model taste = brand / dist=multinomial
                         link=cumlogit
                         aggregate=brand
                         type1
                         ;
   estimate 'LogOR12' brand 1 -1 / exp;
   estimate 'LogOR13' brand 1  0  -1 / exp;
   estimate 'LogOR23' brand 0  1  -1 / exp;
run;
```

The AGGREGATE=BRAND option in the MODEL statement specifies the variable brand as defining multinomial populations for computing deviances and Pearson chi-squares. The RORDER=DATA option specifies that the taste variable levels be ordered by their order of appearance in the input data set, that is, from very good (vg) to very bad (vb). By default, the response is sorted in increasing ASCII order. Always check the "Response Profiles" table to verify that response levels are appropriately ordered. The TYPE1 option requests a Type 1 test for the significance of the covariate brand.

If $\gamma_j(\mathbf{x}) = \Pr(\text{taste} \leq j)$ is the cumulative probability of the $j$th or lower taste category, then the odds ratio comparing $\mathbf{x}_1$ to $\mathbf{x}_2$ is as follows:

$$\frac{\gamma_j(\mathbf{x}_1)/(1-\gamma_j(\mathbf{x}_1))}{\gamma_j(\mathbf{x}_2)/(1-\gamma_j(\mathbf{x}_2))} = \exp[(\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\beta}]$$

Refer to McCullagh and Nelder (1989, Chapter 5) for details on the cumulative logit model. The ESTIMATE statements compute log odds ratios comparing each of brands. The EXP option in the ESTIMATE statements exponentiates the log odds ratios to form odds ratio estimates. Standard errors and confidence intervals are also computed.

Output 29.4.1 displays general information about the model and data, the levels of the CLASS variable brand, and the total number of occurrences of the ordered levels of the response variable taste.

**Output 29.4.1.** Ordinal Model Information

The GENMOD Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.ICECREAM |
| Distribution | Multinomial |
| Link Function | Cumulative Logit |
| Dependent Variable | taste |
| Frequency Weight Variable | count |
| Observations Used | 15 |
| Sum Of Frequency Weights | 1045 |

Class Level Information

| Class | Levels | Values |
|---|---|---|
| brand | 3 | ice1 ice2 ice3 |

Response Profile

| Ordered Level | Ordered Value | Count |
|---|---|---|
| 1 | vg | 140 |
| 2 | g | 162 |
| 3 | m | 421 |
| 4 | b | 156 |
| 5 | vb | 166 |

Output 29.4.2 displays estimates of the intercept terms and covariates and associated statistics. The intercept terms correspond to the four cumulative logits defined on the taste categories in the order shown in Output 29.4.1. That is, Intercept1 is the intercept for the first cumulative logit, $\log(\frac{p_1}{1-p_1})$, Intercept2 is the intercept for the second cumulative logit $\log(\frac{p_1+p_2}{1-(p_1+p_2)})$, and so forth.

**Output 29.4.2.** Parameter Estimates

The GENMOD Procedure

Analysis Of Parameter Estimates

| Parameter | | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square |
|---|---|---|---|---|---|---|---|
| Intercept1 | | 1 | -1.8578 | 0.1219 | -2.0967 | -1.6189 | 232.35 |
| Intercept2 | | 1 | -0.8646 | 0.1056 | -1.0716 | -0.6576 | 67.02 |
| Intercept3 | | 1 | 0.9231 | 0.1060 | 0.7154 | 1.1308 | 75.87 |
| Intercept4 | | 1 | 1.8078 | 0.1191 | 1.5743 | 2.0413 | 230.32 |
| brand | ice1 | 1 | 0.3847 | 0.1370 | 0.1162 | 0.6532 | 7.89 |
| brand | ice2 | 1 | -0.6457 | 0.1397 | -0.9196 | -0.3719 | 21.36 |
| brand | ice3 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Scale | | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | |

Analysis Of Parameter Estimates

| Parameter | | Pr > ChiSq |
|---|---|---|
| Intercept1 | | <.0001 |
| Intercept2 | | <.0001 |
| Intercept3 | | <.0001 |
| Intercept4 | | <.0001 |
| brand | ice1 | 0.0050 |
| brand | ice2 | <.0001 |
| brand | ice3 | . |
| Scale | | |

NOTE: The scale parameter was held fixed.

The Type 1 test displayed in Output 29.4.3 indicates that Brand is highly significant; that is, there are significant differences in the brands. The log odds ratios and odds ratios in the "ESTIMATE Statement Results" table indicate the relative differences between the brands. For example, the odds ratio of 2.8 in the "Exp(LogOR12)" row indicates that the odds of brand 1 being in lower taste categories is 2.8 times the odds of brand 2 being in lower taste categories. Since, in this ordering, the lower categories represent the more favorable taste results, this indicates that brand 1 scored significantly better than brand 2. This is also apparent from the data in this example.

**Output 29.4.3.** Type 1 Tests and Odds Ratios

The GENMOD Procedure

LR Statistics For Type 1 Analysis

| Source | Deviance | DF | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Intercepts | 65.9576 | | | |
| brand | 9.8654 | 2 | 56.09 | <.0001 |

Contrast Estimate Results

| Label | Estimate | Standard Error | Alpha | Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| LogOR12 | 1.0305 | 0.1401 | 0.05 | 0.7559 | 1.3050 | 54.11 | <.0001 |
| Exp(LogOR12) | 2.8024 | 0.3926 | 0.05 | 2.1295 | 3.6878 | | |
| LogOR13 | 0.3847 | 0.1370 | 0.05 | 0.1162 | 0.6532 | 7.89 | 0.0050 |
| Exp(LogOR13) | 1.4692 | 0.2013 | 0.05 | 1.1233 | 1.9217 | | |
| LogOR23 | -0.6457 | 0.1397 | 0.05 | -0.9196 | -0.3719 | 21.36 | <.0001 |
| Exp(LogOR23) | 0.5243 | 0.0733 | 0.05 | 0.3987 | 0.6894 | | |

## Example 29.5. GEE for Binary Data with Logit Link Function

Table 29.4 displays a partial listing of a SAS data set of clinical trial data comparing two treatments for a respiratory disorder. See "Gee Model for Binary Data" in the SAS/STAT Sample Program Library for the complete data set. These data are from Stokes, Davis, and Koch (1995), where a SAS macro is used to fit a GEE model. A GEE model is fit, using the REPEATED statement in the GENMOD procedure.

**Table 29.4.** Respiratory Disorder Data

| Obs | center | id | age | baseline | active | center2 | female | visit | outcome |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 46 | 0 | 0 | 0 | 0 | 1 | 0 |
| 2 | 1 | 1 | 46 | 0 | 0 | 0 | 0 | 2 | 0 |
| 3 | 1 | 1 | 46 | 0 | 0 | 0 | 0 | 3 | 0 |
| 4 | 1 | 1 | 46 | 0 | 0 | 0 | 0 | 4 | 0 |
| 5 | 1 | 2 | 28 | 0 | 0 | 0 | 0 | 1 | 0 |
| 6 | 1 | 2 | 28 | 0 | 0 | 0 | 0 | 2 | 0 |
| 7 | 1 | 2 | 28 | 0 | 0 | 0 | 0 | 3 | 0 |
| 8 | 1 | 2 | 28 | 0 | 0 | 0 | 0 | 4 | 0 |

Patients in each of two centers are randomly assigned to groups receiving the active treatment or a placebo. During treatment, respiratory status (coded here as 0=poor, 1=good) is determined for each of four visits. The variables center, treatment, sex, and baseline (baseline respiratory status) are classification variables with two levels. The variable age (age at time of entry into the study) is a continuous variable.

Explanatory variables in the model are Intercept ($x_{ij1}$), treatment ($x_{ij2}$), center ($x_{ij3}$), sex ($x_{ij4}$), age ($x_{ij6}$), and baseline ($x_{ij6}$), so that $\mathbf{x}_{ij}' = [x_{ij1}, x_{ij2}, \ldots, x_{ij6}]$ is the vector of explanatory variables. Indicator variables for the classification explanatory variables can be automatically generated by listing them in the CLASS statement in PROC GENMOD. However, in order to be consistent with the analysis

in Stokes, Davis, and Koch (1995), the four classification explanatory variables are coded as follows:

$$x_{ij2} = \begin{cases} 0 \text{ placebo} \\ 1 \text{ active} \end{cases} \qquad x_{ij3} = \begin{cases} 0 \text{ center 1} \\ 1 \text{ center 2} \end{cases}$$

$$x_{ij4} = \begin{cases} 0 \text{ male} \\ 1 \text{ female} \end{cases} \qquad x_{ij6} = \begin{cases} 0 \text{ poor} \\ 1 \text{ good} \end{cases}$$

Suppose $y_{ij}$ represents the respiratory status of patient $i$ at the $j$th visit, $j = 1, \ldots, 4$, and $\mu_{ij} = \mathrm{E}(y_{ij})$ represents the mean of the respiratory status. Since the response data are binary, you can use the variance function for the binomial distribution $v(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$ and the logit link function $g(\mu_{ij}) = \log(\mu_{ij}/(1 - \mu_{ij}))$. The model for the mean is $g(\mu_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of regression parameters to be estimated.

Further manipulation of the data set creates an observation for each visit with the respiratory status at each visit represented by the binary variable outcome and indicator variables for treatment (active), center (center2), and sex (female).

```
data resp;
   keep id active center center2 female age baseline visit outcome;
   input center id treatmnt $ sex $ age baseline visit1-visit4;
   active=(treatmnt='A');
   center2=(center=2);
   female=(sex='F');
   visit=1;  outcome=visit1;  output;
   visit=2;  outcome=visit2;  output;
   visit=3;  outcome=visit3;  output;
   visit=4;  outcome=visit4;  output;
   datalines;
1  1 P M 46 0 0 0 0 0
1  2 P M 28 0 0 0 0 0
1  3 A M 23 1 1 1 1 1
1  4 P M 44 1 1 1 1 0
1  5 P F 13 1 1 1 1 1
.
.
.
1 52 P M 43 0 0 0 1 0
1 53 A F 32 0 0 0 1 0
1 54 A M 11 1 1 1 1 0
1 55 P M 24 1 1 1 1 1
1 56 A M 25 0 1 1 0 1
2  1 P F 39 0 0 0 0 0
2  2 A M 25 0 0 1 1 1
2  3 A M 58 1 1 1 1 1
2  4 P F 51 1 1 0 1 1
2  5 P F 32 1 0 0 1 1
```

```
.
.
.
2 51 A M 43 1 1 1 1 0
2 52 A F 39 0 1 1 1 1
2 53 A M 68 0 1 1 1 1
2 54 A F 63 1 1 1 1 1
2 55 A M 31 1 1 1 1 1
;
```

The GEE solution is requested with the REPEATED statement in the GENMOD procedure. The option SUBJECT=ID(CENTER) specifies that the observations in a single cluster are uniquely identified by center and id within center. The option TYPE=UNSTR specifies the unstructured working correlation structure. The MODEL statement specifies the regression model for the mean with the binomial distribution variance function.

```
proc genmod data=resp;
   class id center;
   model outcome=center2 active female age baseline / d=bin;
   repeated  subject=id(center) / type=unstr corrw;
run;
```

These statements first produce the usual output (not shown) for fitting the generalized linear (GLM) model specified in the MODEL statement. The parameter estimates from the GLM model are used as initial values for the GEE solution.

Information about the GEE model is displayed in Output 29.5.1. The results of GEE model fitting are displayed in Output 29.5.2. If you specify no other options, the standard errors, confidence intervals, $Z$ scores, and $p$-values are based on empirical standard error estimates. You can specify the MODELSE option in the REPEATED statement to create a table based on model-based standard error estimates.

**Output 29.5.1.** Model Fitting Information

The GENMOD Procedure

GEE Model Information

| | |
|---|---|
| Correlation Structure | Unstructured |
| Subject Effect | id(center) (111 levels) |
| Number of Clusters | 111 |
| Correlation Matrix Dimension | 4 |
| Maximum Cluster Size | 4 |
| Minimum Cluster Size | 4 |

**Output 29.5.2.** Results of Model Fitting

The GENMOD Procedure

Working Correlation Matrix

| | Col1 | Col2 | Col3 | Col4 |
|---|---|---|---|---|
| Row1 | 1.0000 | 0.3351 | 0.2140 | 0.2953 |
| Row2 | 0.3351 | 1.0000 | 0.4429 | 0.3581 |
| Row3 | 0.2140 | 0.4429 | 1.0000 | 0.3964 |
| Row4 | 0.2953 | 0.3581 | 0.3964 | 1.0000 |

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

| Parameter | Estimate | Standard Error | 95% Confidence Limits | | Z | Pr > \|Z\| |
|---|---|---|---|---|---|---|
| Intercept | -0.8882 | 0.4568 | -1.7835 | 0.0071 | -1.94 | 0.0519 |
| center2 | 0.6558 | 0.3512 | -0.0326 | 1.3442 | 1.87 | 0.0619 |
| active | 1.2442 | 0.3455 | 0.5669 | 1.9214 | 3.60 | 0.0003 |
| female | 0.1128 | 0.4408 | -0.7512 | 0.9768 | 0.26 | 0.7981 |
| age | -0.0175 | 0.0129 | -0.0427 | 0.0077 | -1.36 | 0.1728 |
| baseline | 1.8981 | 0.3441 | 1.2237 | 2.5725 | 5.52 | <.0001 |

The non-significance of age and female make them candidates for omission from the model.

## Example 29.6. Log Odds Ratios and the ALR Algorithm

Since the respiratory data in Example 29.5 are binary, you can use the ALR algorithm to model the log odds ratios instead of using working correlations to model associations. Here, a "fully parameterized cluster" model for the log odds ratio is fit. That is, there is a log odds ratio parameter for each unique pair of responses within clusters, and all clusters are parameterized identically. The following statements fit the same regression model for the mean as in Example 29.5 but use a regression model for the log odds ratios instead of a working correlation. The LOGOR=FULLCLUST option specifies a fully parameterized log odds ratio model.

```
proc genmod data=resp;
   class id center;
   model outcome=center2 active female age baseline / dist=bin;
   repeated  subject=id(center) / logor=fullclust;
run;
```

The results of fitting the model are displayed in Output 29.6.1 along with a table that shows the correspondence between the log odds ratio parameters and the within cluster pairs.

**Output 29.6.1.** Results of Model Fitting

The GENMOD Procedure

Log Odds Ratio Parameter Information

| Parameter | Group |
|---|---|
| Alpha1 | (1, 2) |
| Alpha2 | (1, 3) |
| Alpha3 | (1, 4) |
| Alpha4 | (2, 3) |
| Alpha5 | (2, 4) |
| Alpha6 | (3, 4) |

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

| Parameter | Estimate | Standard Error | 95% Confidence Limits | | Z | Pr > \|Z\| |
|---|---|---|---|---|---|---|
| Intercept | -0.9266 | 0.4513 | -1.8111 | -0.0421 | -2.05 | 0.0400 |
| center2 | 0.6287 | 0.3486 | -0.0545 | 1.3119 | 1.80 | 0.0713 |
| active | 1.2611 | 0.3406 | 0.5934 | 1.9287 | 3.70 | 0.0002 |
| female | 0.1024 | 0.4362 | -0.7526 | 0.9575 | 0.23 | 0.8144 |
| age | -0.0162 | 0.0125 | -0.0407 | 0.0084 | -1.29 | 0.1977 |
| baseline | 1.8980 | 0.3404 | 1.2308 | 2.5652 | 5.58 | <.0001 |
| Alpha1 | 1.6109 | 0.4892 | 0.6522 | 2.5696 | 3.29 | 0.0010 |
| Alpha2 | 1.0771 | 0.4834 | 0.1297 | 2.0246 | 2.23 | 0.0259 |
| Alpha3 | 1.5875 | 0.4735 | 0.6594 | 2.5155 | 3.35 | 0.0008 |
| Alpha4 | 2.1224 | 0.5022 | 1.1381 | 3.1068 | 4.23 | <.0001 |
| Alpha5 | 1.8818 | 0.4686 | 0.9634 | 2.8001 | 4.02 | <.0001 |
| Alpha6 | 2.1046 | 0.4949 | 1.1347 | 3.0745 | 4.25 | <.0001 |

You can fit the same model by fully specifying the *z*-matrix. The following statements create a data set containing the full *z*-matrix.

```
data zin;
   keep id center z1-z6 y1 y2;
   array zin(6) z1-z6;
   set resp ;
   by center id;
   if first.id
      then do;
         t = 0;
         do m = 1 to 4;
            do n = m+1 to 4;
               do j = 1 to 6;
                  zin(j) = 0;
               end;
            y1 = m;
            y2 = n;
            t + 1;
            zin(t) = 1;
            output;
            end;
         end;
      end;
```

```
run;
proc print data=zin (obs=12);
run;
```

Output 29.6.2 displays the full *z*-matrix for the first two clusters. The *z*-matrix is identical for all clusters in this example.

**Output 29.6.2.** Full *z*-Matrix Data Set

| Obs | z1 | z2 | z3 | z4 | z5 | z6 | center | id | y1 | y2 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 2 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 3 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 4 |
| 4 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 2 | 3 |
| 5 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 2 | 4 |
| 6 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 3 | 4 |
| 7 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 2 |
| 8 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 3 |
| 9 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 2 | 1 | 4 |
| 10 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 3 |
| 11 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 2 | 4 |
| 12 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 3 | 4 |

The following statements fit the model for fully parameterized clusters by fully specifying the *z*-matrix. The results are identical to those shown previously.

```
proc genmod data=resp;
   class id center;
   model outcome=center2 active female age baseline / dist=bin;
   repeated  subject=id(center) / logor=zfull
                                  zdata=zin
                                  zrow =(z1-z6)
                                  ypair=(y1 y2) ;
run;
```

## Example 29.7. Log-Linear Model for Count Data

These data, from Thall and Vail (1990), are concerned with the treatment of people suffering from epileptic seizure episodes. These data are also analyzed in Diggle, Liang, and Zeger (1994). The data consist of the number of epileptic seizures in an eight-week baseline period, before any treatment, and in each of four two-week treatment periods, in which patients received either a placebo or the drug Progabide in addition to other therapy. A portion of the data is displayed in Table 29.5. See "Gee Model for Count Data, Exchangeable Correlation" in the SAS/STAT Sample Program Library for the complete data set.

**Table 29.5.** Epileptic Seizure Data

| Patient ID | Treatment | Baseline | Visit1 | Visit2 | Visit3 | Visit4 |
|---|---|---|---|---|---|---|
| 104 | Placebo | 11 | 5 | 3 | 3 | 3 |
| 106 | Placebo | 11 | 3 | 5 | 3 | 3 |
| 107 | Placebo | 6 | 2 | 4 | 0 | 5 |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |
| 101 | Progabide | 76 | 11 | 14 | 9 | 8 |
| 102 | Progabide | 38 | 8 | 7 | 9 | 4 |
| 103 | Progabide | 19 | 0 | 4 | 3 | 0 |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |

Model the data as a log-linear model with $V(\mu) = \mu$ (the Poisson variance function) and

$$\begin{aligned} \log(E(Y_{ij})) &= \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \\ &\quad x_{i1}x_{i2}\beta_3 + \log(t_{ij}) \end{aligned}$$

where

- $Y_{ij}$ = number of epileptic seizures in interval $j$
- $t_{ij}$ = length of interval $j$
- $x_{i1} = \begin{cases} 1: & \text{weeks } 8-16 \text{ (treatment)} \\ 0: & \text{weeks } 0-8 \text{ (baseline)} \end{cases}$
- $x_{i2} = \begin{cases} 1: & \text{progabide group} \\ 0: & \text{placebo group} \end{cases}$

The correlations between the counts are modeled as $r_{ij} = \alpha$, $i \neq j$ (exchangeable correlations). For comparison, the correlations are also modeled as independent (identity correlation matrix). In this model, the regression parameters have the interpretation in terms of the log seizure rate displayed in Table 29.6.

**Table 29.6.** Interpretation of Regression Parameters

| Treatment | Visit | $\log(E(Y_{ij})/t_{ij})$ |
|---|---|---|
| Placebo | Baseline | $\beta_0$ |
| | 1-4 | $\beta_0 + \beta_1$ |
| Progabide | Baseline | $\beta_0 + \beta_2$ |
| | 1-4 | $\beta_0 + \beta_1 + \beta_2 + \beta_3$ |

The difference between the log seizure rates in the pretreatment (baseline) period and the treatment periods is $\beta_1$ for the placebo group and $\beta_1 + \beta_3$ for the Progabide group. A value of $\beta_3 < 0$ indicates a reduction in the seizure rate.

The following statements input the data, which are arranged as one visit per observation:

```
data thall;
   input id y visit trt bline age;
datalines;
104 5 1  0 11 31
104 3 2  0 11 31
104 3 3  0 11 31
104 3 4  0 11 31
106 3 1  0 11 30
106 5 2  0 11 30
106 3 3  0 11 30
106 3 4  0 11 30
107 2 1  0 6 25
107 4 2  0 6 25
107 0 3  0 6 25
107 5 4  0 6 25
114 4 1  0 8 36
114 4 2  0 8 36
...
run;
```

Some further data manipulations create an observation for the baseline measures, a log time interval variable for use as an offset, and an indicator variable for whether the observation is for a baseline measurement or a visit measurement. Patient 207 is deleted as an outlier, as in the Diggle, Liang, and Zeger (1994) analysis.

```
data new;
   set thall;
   output;
   if visit=1 then do;
      y=bline;
      visit=0;
      output;
   end;
run;

data new2;
   set new;
   if id ne 207;
   if visit=0 then do;
      x1=0;
      ltime=log(8);
   end;
   else do;
      x1=1;
      ltime=log(2);
   end;
run;
```

The GEE solution is requested by using the REPEATED statement in the GENMOD procedure. The SUBJECT=ID option indicates that the id variable describes the ob-

servations for a single cluster, and the CORRW option displays the working correlation matrix. The TYPE= option specifies the correlation structure; the value EXCH indicates the exchangeable structure.

```
proc genmod data=new2;
   class id;
   model y=x1 | trt / d=poisson offset=ltime;
   repeated subject=id / corrw covb type=exch;
run;
```

These statements first produce the usual output from fitting a generalized linear model (GLM) to these data. The estimates are used as initial values for the GEE solution.

Information about the GEE model is displayed in Output 29.7.2. The results of fitting the model are displayed in Output 29.7.3. Compare these with the model of independence displayed in Output 29.7.1. The parameter estimates are nearly identical, but the standard errors for the independence case are underestimated. The coefficient of the interaction term, $\beta_3$, is highly significant under the independence model and marginally significant with the exchangeable correlations model.

**Output 29.7.1.** Independence Model

The GENMOD Procedure

Analysis Of Initial Parameter Estimates

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 1.3476 | 0.0341 | 1.2809 | 1.4144 | 1565.44 | <.0001 |
| x1 | 1 | 0.1108 | 0.0469 | 0.0189 | 0.2027 | 5.58 | 0.0181 |
| trt | 1 | -0.1080 | 0.0486 | -0.2034 | -0.0127 | 4.93 | 0.0264 |
| x1*trt | 1 | -0.3016 | 0.0697 | -0.4383 | -0.1649 | 18.70 | <.0001 |
| Scale | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

NOTE: The scale parameter was held fixed.

**Output 29.7.2.** GEE Model Information

The GENMOD Procedure

GEE Model Information

| | |
|---|---|
| Correlation Structure | Exchangeable |
| Subject Effect | id (58 levels) |
| Number of Clusters | 58 |
| Correlation Matrix Dimension | 5 |
| Maximum Cluster Size | 5 |
| Minimum Cluster Size | 5 |

**Output 29.7.3.** GEE Parameter Estimates

The GENMOD Procedure

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

| Parameter | Estimate | Standard Error | 95% Confidence Limits | | Z | Pr > \|Z\| |
|---|---|---|---|---|---|---|
| Intercept | 1.3476 | 0.1574 | 1.0392 | 1.6560 | 8.56 | <.0001 |
| x1 | 0.1108 | 0.1161 | -0.1168 | 0.3383 | 0.95 | 0.3399 |
| trt | -0.1080 | 0.1937 | -0.4876 | 0.2716 | -0.56 | 0.5770 |
| x1*trt | -0.3016 | 0.1712 | -0.6371 | 0.0339 | -1.76 | 0.0781 |

Table 29.7 displays the regression coefficients, standard errors, and normalized coefficients that result from fitting the model using independent and exchangeable working correlation matrices.

**Table 29.7.** Results of Model Fitting

| Variable | Correlation Structure | Coef. | Std. Error | Coef./S.E. |
|---|---|---|---|---|
| Intercept | Exchangeable | 1.35 | 0.16 | 8.56 |
| | Independent | 1.35 | 0.03 | 39.52 |
| Visit $(x_1)$ | Exchangeable | 0.11 | 0.12 | 0.95 |
| | Independent | 0.11 | 0.05 | 2.36 |
| Treat $(x_2)$ | Exchangeable | $-0.11$ | 0.19 | $-0.56$ |
| | Independent | $-0.11$ | 0.05 | $-2.22$ |
| $x_1 * x_2$ | Exchangeable | $-0.30$ | 0.17 | $-1.76$ |
| | Independent | $-0.30$ | 0.07 | $-4.32$ |

The fitted exchangeable correlation matrix is specified with the CORRW option and is displayed in Output 29.7.4.

**Output 29.7.4.** Working Correlation Matrix

The GENMOD Procedure

Working Correlation Matrix

| | Col1 | Col2 | Col3 | Col4 | Col5 |
|---|---|---|---|---|---|
| Row1 | 1.0000 | 0.5941 | 0.5941 | 0.5941 | 0.5941 |
| Row2 | 0.5941 | 1.0000 | 0.5941 | 0.5941 | 0.5941 |
| Row3 | 0.5941 | 0.5941 | 1.0000 | 0.5941 | 0.5941 |
| Row4 | 0.5941 | 0.5941 | 0.5941 | 1.0000 | 0.5941 |
| Row5 | 0.5941 | 0.5941 | 0.5941 | 0.5941 | 1.0000 |

If you specify the COVB option, you produce both the model-based (naive) and the empirical (robust) covariance matrices. Output 29.7.5 contains these estimates.

**Output 29.7.5.** Covariance Matrices

The GENMOD Procedure

Covariance Matrix (Model-Based)

| | Prm1 | Prm2 | Prm3 | Prm4 |
|---|---|---|---|---|
| Prm1 | 0.01223 | 0.001520 | -0.01223 | -0.001520 |
| Prm2 | 0.001520 | 0.01519 | -0.001520 | -0.01519 |
| Prm3 | -0.01223 | -0.001520 | 0.02495 | 0.005427 |
| Prm4 | -0.001520 | -0.01519 | 0.005427 | 0.03748 |

Covariance Matrix (Empirical)

| | Prm1 | Prm2 | Prm3 | Prm4 |
|---|---|---|---|---|
| Prm1 | 0.02476 | -0.001152 | -0.02476 | 0.001152 |
| Prm2 | -0.001152 | 0.01348 | 0.001152 | -0.01348 |
| Prm3 | -0.02476 | 0.001152 | 0.03751 | -0.002999 |
| Prm4 | 0.001152 | -0.01348 | -0.002999 | 0.02931 |

The two covariance estimates are similar, indicating an adequate correlation model.

# References


Agresti, A. (1990), *Categorical Data Analysis,* New York: John Wiley & Sons, Inc.

Aitkin, M. (1987), "Modelling Variance Heterogeneity in Normal Regression Using GLIM," *Applied Statistics*, 36, 332–339.

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

Boos, D. (1992), "On Generalized Score Tests," *The American Statistician*, 46, 327–333.

Carey, V., Zeger, S.L., and Diggle, P. (1993), "Modelling Multivariate Binary Data with Alternating Logistic Regressions," *Biometrika*, 80, 517–526.

Davison, A.C. and Snell, E.J. (1991), "Residuals and Diagnostics," in *Statistical Theory and Modelling,* eds. D.V. Hinkley, N. Reid, and E.J. Snell, London: Chapman and Hall.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994), *Analysis of Longitudinal Data*, Oxford: Clarendon Press.

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.

Hilbe, J. (1994), "Log Negative Binomial Regression Using the GENMOD Procedure," *Proceedings of the Nineteenth Annual SAS User's Group International Conference,* 14, 1199–1204

Jennrich, R.I. and Schluchter, M.D. (1986), "Unbalanced Repeated-Measures Models with Structured Covariance Matrices," *Biometrics*, 42, 805–820.

Lawless, J.E. (1982), *Statistical Models And Methods For Lifetime Data*, New York: John Wiley & Sons, Inc.

Lawless, J.E. (1987), "Negative Binomial and Mixed Poisson Regression," *The Canadian Journal of Statistics*, 15, 209–225.

Liang, K.Y. and Zeger, S.L. (1986), "Longitudinal Data Analysis using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lipsitz, S.H., Fitzmaurice, G.M., Orav, E.J., and Laird, N.M. (1994), "Performance of Generalized Estimating Equations in Practical Situations," *Biometrics*, 50, 270–278.

Lipsitz, S.H., Kim, K., and Zhao, L. (1994), "Analysis of Repeated Categorical Data Using Generalized Estimating Equations," *Statistics in Medicine*, 13, 1149–1163.

Lipsitz, S.H., Laird, N.M., and Harrington, D.P., (1991), "Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as a Measure of Association," *Biometrika*, 78, 153–160.

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

McCullagh, P. (1983), "Quasi-Likelihood Functions," *Annals of Statistics*, 11, 59–67.

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

Miller, M.E., Davis, C.S., and Landis, J.R. (1993), "The Analysis of Longitudinal Polytomous Data: Generalized Estimating Equations and Connections with Weighted Least Squares," *Biometrics*, 49, 1033–1044.

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

Nelson, W. (1982), *Applied Life Data Analysis*, New York: John Wiley & Sons, Inc.

Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons, Inc.

Rotnitzky, A. and Jewell, N.P., (1990), "Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Cluster Correlated Data," *Biometrika*, 77, 485–497.

Searle, S.R. (1971), *Linear Models*, New York: John Wiley & Sons, Inc.

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

Thall, P.F. and Vail, S.C. (1990), "Some Covariance Models for Longitudinal Count Data with Overdispersion," *Biometrics*, 46, 657–671.

Ware, J.H., Dockery, Spiro A. III, Speizer, F.E., and Ferris, B.G., Jr. (1984), "Passive Smoking, Gas Cooking, and Respiratory Health of Children Living in Six Cities," *American Review of Respiratory Diseases*, 129, 366–374.

Williams, D.A. (1987), "Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions," *Applied Statistics*, 36, 181–191.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

**SAS/STAT® User's Guide, Version 8**



ISBN 1–58025–494–2



SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999