# SAS/STAT® User's Guide

# The GENMOD Procedure

## 2022.12*

* This document might apply to additional versions of the software. Open this document in SAS Help Center and click on the version in the banner to see all available versions.

SAS® Documentation
December 15, 2022

This document is an individual chapter from *SAS/STAT® User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2022. *SAS/STAT® User's Guide*. Cary, NC: SAS Institute Inc.

**SAS/STAT® User's Guide**

December 2022

# Chapter 51
# The GENMOD Procedure

## Contents

Overview: GENMOD Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **3555**

What Is a Generalized Linear Model? . . . . . . . . . . . . . . . . . . . . . . . . . . . 3556

Examples of Generalized Linear Models . . . . . . . . . . . . . . . . . . . . . . . . . . 3557

The GENMOD Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3558

Getting Started: GENMOD Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . **3561**

Poisson Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3561

Bayesian Analysis of a Linear Regression Model . . . . . . . . . . . . . . . . . . . . 3565

Generalized Estimating Equations . . . . . . . . . . . . . . . . . . . . . . . . . . . 3577

Syntax: GENMOD Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **3580**

PROC GENMOD Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3581

ASSESS Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3586

BAYES Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3587

BY Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3597

CLASS Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3598

CODE Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3601

CONTRAST Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3602

DEVIANCE Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3604

EFFECTPLOT Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3605

ESTIMATE Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3606

EXACT Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3607

EXACTOPTIONS Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3610

FREQ Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3613

FWDLINK Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3613

INVLINK Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3613

LSMEANS Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3614

LSMESTIMATE Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3615

MODEL Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3616

OUTPUT Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3627

Programming Statements . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3630

REPEATED Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3632

SLICE Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3637

STORE Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3637

STRATA Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3637

VARIANCE Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3639

WEIGHT Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3639

ZEROMODEL Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3639

Details: GENMOD Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **3640**
Generalized Linear Models Theory . . . . . . . . . . . . . . . . . . . . . . . . . . 3640
Specification of Effects . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3651
Parameterization Used in PROC GENMOD . . . . . . . . . . . . . . . . . . . . . . . 3652
Type 1 Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3653
Type 3 Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3654
Confidence Intervals for Parameters . . . . . . . . . . . . . . . . . . . . . . . . 3655
*F* Statistics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3656
Lagrange Multiplier Statistics . . . . . . . . . . . . . . . . . . . . . . . . . . . 3657
Predicted Values of the Mean . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3657
Residuals . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3658
Multinomial Models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3659
Zero-Inflated Models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3660
Tweedie Distribution for Generalized Linear Models . . . . . . . . . . . . . . . . . 3661
Generalized Estimating Equations . . . . . . . . . . . . . . . . . . . . . . . . . . 3663
Assessment of Models Based on Aggregates of Residuals . . . . . . . . . . . . . . . 3672
Case Deletion Diagnostic Statistics . . . . . . . . . . . . . . . . . . . . . . . . . 3675
Bayesian Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3679
Exact Logistic and Exact Poisson Regression . . . . . . . . . . . . . . . . . . . . 3684
Response Level Ordering . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3687
Missing Values . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3688
Displayed Output for Classical Analysis . . . . . . . . . . . . . . . . . . . . . . 3688
Displayed Output for Bayesian Analysis . . . . . . . . . . . . . . . . . . . . . . . 3696
Displayed Output for Exact Analysis . . . . . . . . . . . . . . . . . . . . . . . . 3698
ODS Table Names . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3699
ODS Graphics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3703
Examples: GENMOD Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **3705**
Example 51.1: Logistic Regression . . . . . . . . . . . . . . . . . . . . . . . . . 3705
Example 51.2: Normal Regression, Log Link . . . . . . . . . . . . . . . . . . . . . 3707
Example 51.3: Gamma Distribution Applied to Life Data . . . . . . . . . . . . . . . 3709
Example 51.4: Ordinal Model for Multinomial Data . . . . . . . . . . . . . . . . . . 3712
Example 51.5: GEE for Binary Data with Logit Link Function . . . . . . . . . . . . . 3715
Example 51.6: Log Odds Ratios and the ALR Algorithm . . . . . . . . . . . . . . . . 3718
Example 51.7: Log-Linear Model for Count Data . . . . . . . . . . . . . . . . . . . 3721
Example 51.8: Model Assessment of Multiple Regression Using Aggregates of Residuals 3725
Example 51.9: Assessment of a Marginal Model for Dependent Data . . . . . . . . . 3733
Example 51.10: Bayesian Analysis of a Poisson Regression Model . . . . . . . . . . 3736
Example 51.11: Exact Poisson Regression . . . . . . . . . . . . . . . . . . . . . . 3750
Example 51.12: Tweedie Regression . . . . . . . . . . . . . . . . . . . . . . . . . 3753
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **3755**

# Overview: GENMOD Procedure

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.

See McCullagh and Nelder (1989) for a discussion of statistical modeling using generalized linear models. The books by Aitkin et al. (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. Myers, Montgomery, and Vining (2002) provide applications of generalized linear models in the engineering and physical sciences. Collett (2003) and Hilbe (2009) provide comprehensive accounts of generalized linear models when the responses are binary.

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 might 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) by using GEEs. For more details on GEEs, see Hardin and Hilbe (2003); Diggle, Liang, and Zeger (1994); Lipsitz et al. (1994).

Bayesian analysis of generalized linear models can be requested by using the BAYES statement in the GENMOD procedure. In Bayesian analysis, the model parameters are treated as random variables, and inference about parameters is based on the posterior distribution of the parameters, given the data. The posterior distribution is obtained using Bayes' theorem as the likelihood function of the data weighted with a prior distribution. The prior distribution enables you to incorporate knowledge or experience of the likely range of values of the parameters of interest into the analysis. If you have no prior knowledge of the parameter values, you can use a noninformative prior distribution, and the results of the Bayesian analysis will be very similar to a classical analysis based on maximum likelihood. A closed form of the posterior distribution is often not feasible, and a Markov chain Monte Carlo method by Gibbs sampling is used to simulate samples from the posterior distribution. See Chapter 8, "Introduction to Bayesian Analysis Procedures," for an introduction to the basic concepts of Bayesian statistics. Also see the section "Bayesian Analysis: Advantages and Disadvantages" on page 154 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for a discussion of the advantages and disadvantages of Bayesian analysis. See Ibrahim, Chen, and Sinha (2001) for a detailed description of Bayesian analysis.

In a Bayesian analysis, a Gibbs chain of samples from the posterior distribution is generated for the model parameters. Summary statistics (mean, standard deviation, quartiles, HPD and credible intervals, correlation matrix) and convergence diagnostics (autocorrelations; Gelman-Rubin, Geweke, Raftery-Lewis, and Heidelberger and Welch tests; the effective sample size; and Monte Carlo standard errors) are computed for each parameter, as well as the correlation matrix and the covariance matrix of the posterior sample. Trace plots, posterior density plots, and autocorrelation function plots that are created using ODS Graphics are also provided for each parameter.

The GENMOD procedure enables you to perform exact logistic regression, also called exact conditional binary logistic regression, and exact Poisson regression, also called exact conditional Poisson regression, by specifying one or more EXACT statements. You can test individual parameters or conduct a joint test for several parameters. The procedure computes two exact tests: the exact conditional score test and the exact conditional probability test. You can request exact estimation of specific parameters and corresponding odds ratios where appropriate. Point estimates, standard errors, and confidence intervals are provided.

The GENMOD procedure uses ODS Graphics to create graphs as part of its output. For general information about ODS Graphics, see Chapter 24, "Statistical Graphics Using ODS."

## 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 such as the following for which they are not appropriate.

- It might not be reasonable to assume that data are normally distributed. For example, the normal distribution (which is continuous) might 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 might 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 might 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 $i = 1, 2, \ldots$ 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* $V$:

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

  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 must be estimated.

See the section "Response Probability Distributions" on page 3640 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 by 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, $g(\mu) = \mu$

### Logistic Regression

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

### Poisson Regression in Log-Linear Model

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

### Gamma Model with Log Link

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

## 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 as follows:

- 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 as follows:

- 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: $V(\mu) = \mu + k\mu^2$
- geometric: $V(\mu) = \mu + \mu^2$
- multinomial
- zero-inflated Poisson
- zero-inflated negative binomial

The negative binomial and zero-inflated negative binomial are distributions 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. For discussions of the negative binomial distribution, see McCullagh and Nelder (1989); Hilbe (1994, 2007); Long (1997); Cameron and Trivedi (1998); Lawless (1987).

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 3659 and McCullagh and Nelder (1989, Chapter 5) for a description of the multinomial distribution.

The zero-inflated Poisson and zero-inflated negative binomial are included in PROC GENMOD even though they are not generalized linear models. They are useful extensions of generalized linear models. See the section "Zero-Inflated Models" on page 3660 for information about the zero-inflated distributions. Models for data with correlated responses fit by the GEE method are not available for zero-inflated distributions.

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 3647 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 to 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 might 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 53, "The GLM Procedure," and Chapter 16, "The Four Types of Estimable Functions," for more information about Type III estimable functions. Also see Littell, Freund, and Spector (1991).

Additional features of the GENMOD procedure include the following:

- 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 51.13 and Table 51.14)
- 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
- ability to perform Bayesian analysis by Gibbs sampling

# Getting Started: GENMOD Procedure

## 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 et al. (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 *log-linear 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

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

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

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

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 that 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, is observed. PROC GENMOD produces the following default output from the preceding statements.

**Figure 51.1** Model Information

**The GENMOD Procedure**

| Model Information | |
|---|---|
| **Data Set** | WORK.INSURE |
| **Distribution** | Poisson |
| **Link Function** | Log |
| **Dependent Variable** | c |
| **Offset Variable** | ln |

The “Model Information” table displayed in Figure 51.1 provides information about the specified model and the input data set.

**Figure 51.2** Class Level Information

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **car** | 3 | large medium small |
| **age** | 2 | 1 2 |

Figure 51.2 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 character variable, and the values are sorted in alphabetical order. This is the default sort order, but you can select different sort orders with the ORDER= option in the PROC GENMOD statement.

**Figure 51.3** Goodness of Fit

| 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 | |
| **Full Log Likelihood** | | -16.4638 | |
| **AIC (smaller is better)** | | 40.9276 | |
| **AICC (smaller is better)** | | 80.9276 | |
| **BIC (smaller is better)** | | 40.0946 | |

The “Criteria For Assessing Goodness Of Fit” table displayed in Figure 51.3 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.

**Figure 51.4** Analysis of Parameter Estimates

| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Wald 95% | | | |
| | | | | Standard | Confidence | | Wald | |
| **Parameter** | | **DF** | **Estimate** | **Error** | **Limits** | | **Chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -1.3168 | 0.0903 | -1.4937 | -1.1398 | 212.73 | <.0001 |
| **car** | large | 1 | -1.7643 | 0.2724 | -2.2981 | -1.2304 | 41.96 | <.0001 |
| **car** | medium | 1 | -0.6928 | 0.1282 | -0.9441 | -0.4414 | 29.18 | <.0001 |
| **car** | small | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **age** | 1 | 1 | -1.3199 | 0.1359 | -1.5863 | -1.0536 | 94.34 | <.0001 |
| **age** | 2 | 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.

Figure 51.4 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 section “Response Probability Distributions” on page 3640 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 3650 for more about 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, as follows:

```
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 figures that follow.

**Figure 51.5** Type 1 Analysis

**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 |

In the table for Type 1 analysis displayed in Figure 51.5, 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.

**Figure 51.6** Type 3 Analysis

**LR Statistics For Type 3 Analysis**

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

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.

## Bayesian Analysis of a Linear Regression Model

Neter et al. (1996) describe a study of 54 patients undergoing a certain kind of liver operation in a surgical unit. The data set Surg contains survival time and certain covariates for each patient. Observations for the first 20 patients in the data set Surg are shown in Figure 51.7.

**Figure 51.7** Surgical Unit Data

| Obs | x1 | x2 | x3 | x4 | y | logy | Logx1 |
|---|---|---|---|---|---|---|---|
| **1** | 6.7 | 62 | 81 | 2.59 | 200 | 2.3010 | 1.90211 |
| **2** | 5.1 | 59 | 66 | 1.70 | 101 | 2.0043 | 1.62924 |
| **3** | 7.4 | 57 | 83 | 2.16 | 204 | 2.3096 | 2.00148 |
| **4** | 6.5 | 73 | 41 | 2.01 | 101 | 2.0043 | 1.87180 |
| **5** | 7.8 | 65 | 115 | 4.30 | 509 | 2.7067 | 2.05412 |
| **6** | 5.8 | 38 | 72 | 1.42 | 80 | 1.9031 | 1.75786 |
| **7** | 5.7 | 46 | 63 | 1.91 | 80 | 1.9031 | 1.74047 |
| **8** | 3.7 | 68 | 81 | 2.57 | 127 | 2.1038 | 1.30833 |
| **9** | 6.0 | 67 | 93 | 2.50 | 202 | 2.3054 | 1.79176 |
| **10** | 3.7 | 76 | 94 | 2.40 | 203 | 2.3075 | 1.30833 |
| **11** | 6.3 | 84 | 83 | 4.13 | 329 | 2.5172 | 1.84055 |
| **12** | 6.7 | 51 | 43 | 1.86 | 65 | 1.8129 | 1.90211 |
| **13** | 5.8 | 96 | 114 | 3.95 | 830 | 2.9191 | 1.75786 |
| **14** | 5.8 | 83 | 88 | 3.95 | 330 | 2.5185 | 1.75786 |
| **15** | 7.7 | 62 | 67 | 3.40 | 168 | 2.2253 | 2.04122 |
| **16** | 7.4 | 74 | 68 | 2.40 | 217 | 2.3365 | 2.00148 |
| **17** | 6.0 | 85 | 28 | 2.98 | 87 | 1.9395 | 1.79176 |
| **18** | 3.7 | 51 | 41 | 1.55 | 34 | 1.5315 | 1.30833 |
| **19** | 7.3 | 68 | 74 | 3.56 | 215 | 2.3324 | 1.98787 |
| **20** | 5.6 | 57 | 87 | 3.02 | 172 | 2.2355 | 1.72277 |

Consider the model

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{LogX1} + \beta_2 \mathrm{X2} + \beta_3 \mathrm{X3} + \beta_4 \mathrm{X4} + \epsilon$$

where Y is the survival time, LogX1 is log(blood-clotting score), X2 is a prognostic index, X3 is an enzyme function test score, X4 is a liver function test score, and $\epsilon$ is an $N(0, \sigma^2)$ error term.

A question of scientific interest is whether blood clotting score has a positive effect on survival time. Using PROC GENMOD, you can obtain a maximum likelihood estimate of the coefficient and construct a null point hypothesis to test whether $\beta_1$ is equal to 0. However, if you are interested in finding the probability that the coefficient is positive, Bayesian analysis offers a convenient alternative. You can use Bayesian analysis to directly estimate the conditional probability, $\Pr(\beta_1 > 0|\mathbf{Y})$, using the posterior distribution samples, which are produced as part of the output by PROC GENMOD.

The example that follows shows how to use PROC GENMOD to carry out a Bayesian analysis of the linear model with a normal error term. The SEED= option is specified to maintain reproducibility; no other options are specified in the BAYES statement. By default, a uniform prior distribution is assumed on the regression coefficients. The uniform prior is a flat prior on the real line with a distribution that reflects ignorance of the location of the parameter, placing equal likelihood on all possible values the regression coefficient can take. Using the uniform prior in the following example, you would expect the Bayesian estimates to resemble the classical results of maximizing the likelihood. If you can elicit an informative prior distribution for the regression coefficients, you should use the COEFFPRIOR= option to specify it. A default noninformative gamma prior is used for the scale parameter $\sigma$.

You should make sure that the posterior distribution samples have achieved convergence before using them for Bayesian inference. PROC GENMOD produces three convergence diagnostics by default. If ODS

Graphics is enabled as specified in the following SAS statements, diagnostic plots are also displayed. See the section “Assessing Markov Chain Convergence” on page 162 in Chapter 8, “Introduction to Bayesian Analysis Procedures,” for more information about convergence diagnostics and their interpretation.

Summary statistics of the posterior distribution samples are produced by default. However, these statistics might not be sufficient for carrying out your Bayesian inference, and further processing of the posterior samples might be necessary. The following SAS statements request the Bayesian analysis, and the OUTPOST= option saves the samples in the SAS data set PostSurg for further processing:

```
proc genmod data=Surg;
   model y = Logx1 X2 X3 X4 / dist=normal;
   bayes seed=1 OutPost=PostSurg;
run;
```

The results of this analysis are shown in the following figures. The “Model Information” table in Figure 51.8 summarizes information about the model you fit and the size of the simulation.

**Figure 51.8** Model Information

**The GENMOD Procedure**

**Bayesian Analysis**

| **Model Information** | | |
|---|---|---|
| **Data Set** | WORK.SURG | |
| **Burn-In Size** | 2000 | |
| **MC Sample Size** | 10000 | |
| **Thinning** | 1 | |
| **Sampling Algorithm** | Conjugate | |
| **Distribution** | Normal | |
| **Link Function** | Identity | |
| **Dependent Variable** | y | Survival Time |

The “Analysis of Maximum Likelihood Parameter Estimates” table in Figure 51.9 summarizes maximum likelihood estimates of the model parameters.

**Figure 51.9** Maximum Likelihood Parameter Estimates

**Analysis Of Maximum Likelihood Parameter Estimates**

| | | | **Standard** | **Wald 95%** | |
|---|---|---|---|---|---|
| **Parameter** | **DF** | **Estimate** | **Error** | **Confidence Limits** | |
| **Intercept** | 1 | -730.559 | 85.4333 | -898.005 | -563.112 |
| **Logx1** | 1 | 171.8758 | 38.2250 | 96.9561 | 246.7954 |
| **x2** | 1 | 4.3019 | 0.5566 | 3.2109 | 5.3929 |
| **x3** | 1 | 4.0309 | 0.4996 | 3.0517 | 5.0100 |
| **x4** | 1 | 18.1377 | 12.0721 | -5.5232 | 41.7986 |
| **Scale** | 1 | 59.8591 | 5.7599 | 49.5705 | 72.2832 |

**Note:** The scale parameter was estimated by maximum likelihood.

Since no prior distributions for the regression coefficients were specified, the default noninformative uniform distributions shown in the "Uniform Prior for Regression Coefficients" table in Figure 51.10 are used. Noninformative priors are appropriate if you have no prior knowledge of the likely range of values of the parameters, and if you want to make probability statements about the parameters or functions of the parameters. See, for example, Ibrahim, Chen, and Sinha (2001) for more information about choosing prior distributions.

**Figure 51.10** Regression Coefficient Priors

**The GENMOD Procedure**

**Bayesian Analysis**

**Uniform Prior for Regression Coefficients**

| Parameter | Prior |
|---|---|
| Intercept | Constant |
| Logx1 | Constant |
| x2 | Constant |
| x3 | Constant |
| x4 | Constant |

The default noninformative improper prior distribution for the normal dispersion parameter is shown in the "Independent Prior Distributions for Model Parameters" table in Figure 51.11.

**Figure 51.11** Scale Parameter Prior

**Independent Prior Distributions for Model Parameters**

| Parameter | Prior Distribution |
|---|---|
| Dispersion | Improper |

By default, the maximum likelihood estimates of the regression parameters are used as the starting values for the simulation when noninformative prior distributions are used. These are listed in the "Initial Values and Seeds" table in Figure 51.12.

**Figure 51.12** MCMC Initial Values and Seeds

**Initial Values of the Chain**

| Chain | Seed | Intercept | Logx1 | x2 | x3 | x4 | Dispersion |
|---|---|---|---|---|---|---|---|
| 1 | 1 | -730.559 | 171.8758 | 4.301896 | 4.030878 | 18.1377 | 3449.176 |

Summary statistics for the posterior sample are displayed in the "Fit Statistics," "Descriptive Statistics for the Posterior Sample," "Interval Statistics for the Posterior Sample," and "Posterior Correlation Matrix" tables in Figure 51.13, Figure 51.14, Figure 51.15, and Figure 51.16, respectively.

**Figure 51.13** Fit Statistics

| Fit Statistics | |
|---|---|
| DIC (smaller is better) | 607.796 |
| pD (effective number of parameters) | 6.062 |

**Figure 51.14** Descriptive Statistics

**The GENMOD Procedure**

**Bayesian Analysis**

**Posterior Summaries**

| | | | | Percentiles | | |
|---|---|---|---|---|---|---|
| **Parameter** | **N** | **Mean** | **Standard Deviation** | **25%** | **50%** | **75%** |
| **Intercept** | 10000 | -730.0 | 91.2102 | -789.6 | -729.6 | -670.5 |
| **Logx1** | 10000 | 171.7 | 40.6455 | 144.2 | 171.6 | 198.6 |
| **x2** | 10000 | 4.2988 | 0.5952 | 3.9029 | 4.2919 | 4.6903 |
| **x3** | 10000 | 4.0308 | 0.5359 | 3.6641 | 4.0267 | 4.3921 |
| **x4** | 10000 | 18.0858 | 12.9123 | 9.4471 | 18.1230 | 26.8141 |
| **Dispersion** | 10000 | 4113.1 | 867.7 | 3497.2 | 3995.9 | 4606.4 |

**Figure 51.15** Interval Statistics

**Posterior Intervals**

| **Parameter** | **Alpha** | **Equal-Tail Interval** | | **HPD Interval** | |
|---|---|---|---|---|---|
| **Intercept** | 0.050 | -908.6 | -549.8 | -906.9 | -549.1 |
| **Logx1** | 0.050 | 91.9723 | 252.5 | 94.1279 | 254.0 |
| **x2** | 0.050 | 3.1091 | 5.4778 | 3.1705 | 5.5167 |
| **x3** | 0.050 | 2.9803 | 5.1031 | 2.9227 | 5.0343 |
| **x4** | 0.050 | -7.3043 | 43.6387 | -8.8440 | 41.8229 |
| **Dispersion** | 0.050 | 2741.5 | 6096.6 | 2540.1 | 5810.0 |

**Figure 51.16** Posterior Sample Correlation Matrix

**Posterior Correlation Matrix**

| **Parameter** | **Intercept** | **Logx1** | **x2** | **x3** | **x4** | **Dispersion** |
|---|---|---|---|---|---|---|
| **Intercept** | 1.000 | -0.857 | -0.579 | -0.712 | 0.582 | 0.000 |
| **Logx1** | -0.857 | 1.000 | 0.286 | 0.491 | -0.640 | 0.007 |
| **x2** | -0.579 | 0.286 | 1.000 | 0.302 | -0.489 | -0.009 |
| **x3** | -0.712 | 0.491 | 0.302 | 1.000 | -0.618 | -0.006 |
| **x4** | 0.582 | -0.640 | -0.489 | -0.618 | 1.000 | 0.003 |
| **Dispersion** | 0.000 | 0.007 | -0.009 | -0.006 | 0.003 | 1.000 |

Since noninformative prior distributions were used, the posterior sample means, standard deviations, and interval statistics shown in Figure 51.13 and Figure 51.14 are consistent with the maximum likelihood estimates shown in Figure 51.9.

By default, PROC GENMOD computes three convergence diagnostics: the lag1, lag5, lag10, and lag50 autocorrelations (Figure 51.17); Geweke diagnostic statistics (Figure 51.18); and effective sample sizes (Figure 51.19). There is no indication that the Markov chain has not converged. See the section "Assessing Markov Chain Convergence" on page 162 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for more information about convergence diagnostics and their interpretation.

**Figure 51.17** Posterior Sample Autocorrelations

**The GENMOD Procedure**

**Bayesian Analysis**

**Posterior Autocorrelations**

| Parameter | Lag 1 | Lag 5 | Lag 10 | Lag 50 |
|---|---|---|---|---|
| **Intercept** | -0.0059 | -0.0037 | -0.0152 | 0.0010 |
| **Logx1** | -0.0002 | -0.0064 | -0.0066 | -0.0054 |
| **x2** | -0.0120 | -0.0026 | -0.0267 | -0.0168 |
| **x3** | 0.0036 | 0.0033 | -0.0035 | 0.0004 |
| **x4** | 0.0034 | -0.0064 | 0.0083 | -0.0124 |
| **Dispersion** | -0.0011 | 0.0091 | -0.0279 | 0.0037 |

**Figure 51.18** Geweke Diagnostic Statistics

**Geweke Diagnostics**

| Parameter | z | Pr > \|z\| |
|---|---|---|
| **Intercept** | -1.0815 | 0.2795 |
| **Logx1** | 1.6667 | 0.0956 |
| **x2** | 0.0977 | 0.9222 |
| **x3** | 0.2506 | 0.8021 |
| **x4** | -1.1082 | 0.2678 |
| **Dispersion** | 0.2451 | 0.8064 |

**Figure 51.19** Effective Sample Sizes

**Effective Sample Sizes**

| | | Autocorrelation | |
|---|---|---|---|
| Parameter | ESS | Time | Efficiency |
| **Intercept** | 10000.0 | 1.0000 | 1.0000 |
| **Logx1** | 10000.0 | 1.0000 | 1.0000 |
| **x2** | 10245.2 | 0.9761 | 1.0245 |
| **x3** | 10000.0 | 1.0000 | 1.0000 |
| **x4** | 10000.0 | 1.0000 | 1.0000 |
| **Dispersion** | 10000.0 | 1.0000 | 1.0000 |

Trace, autocorrelation, and density plots for the seven model parameters, shown in Figure 51.20 through Figure 51.25, are useful in diagnosing whether the Markov chain of posterior samples has converged. These plots show no evidence that the chain has not converged. See the section "Visual Analysis via Trace Plots" on page 163 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for help with interpreting these diagnostic plots.

**Figure 51.20** Diagnostic Plots for Intercept

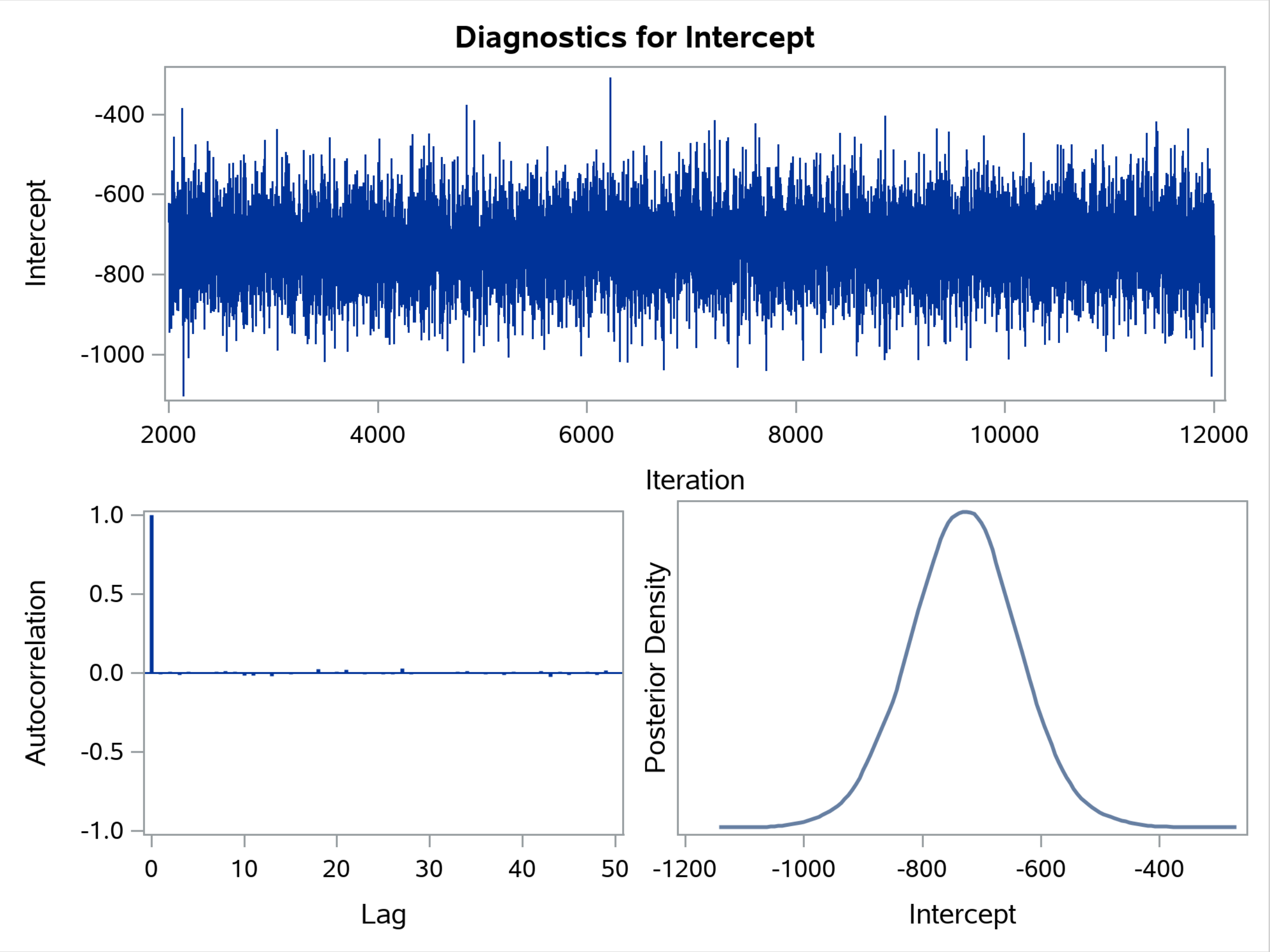


**Figure 51.21** Diagnostic Plots for logX1

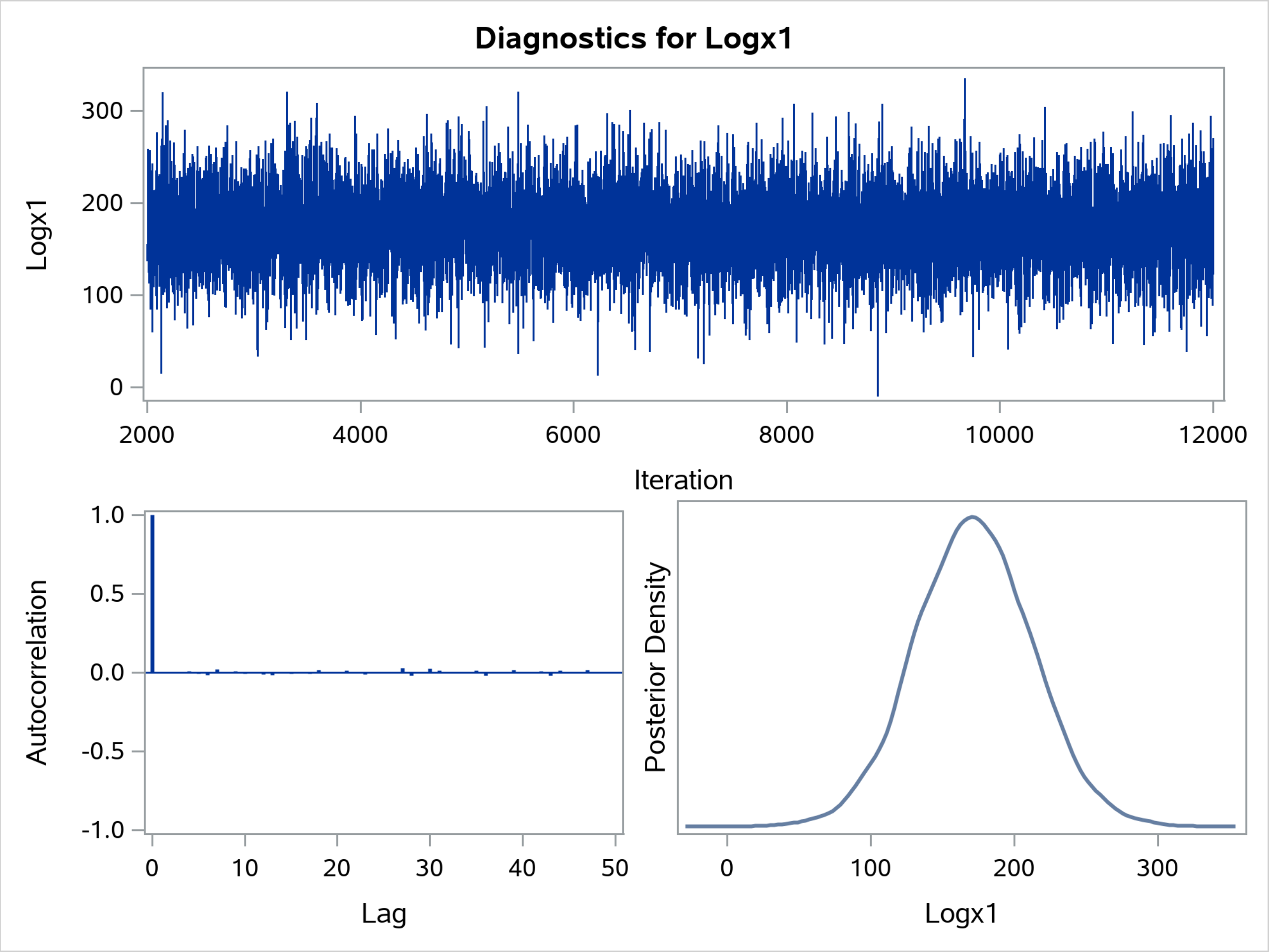


**Figure 51.22** Diagnostic Plots for X2

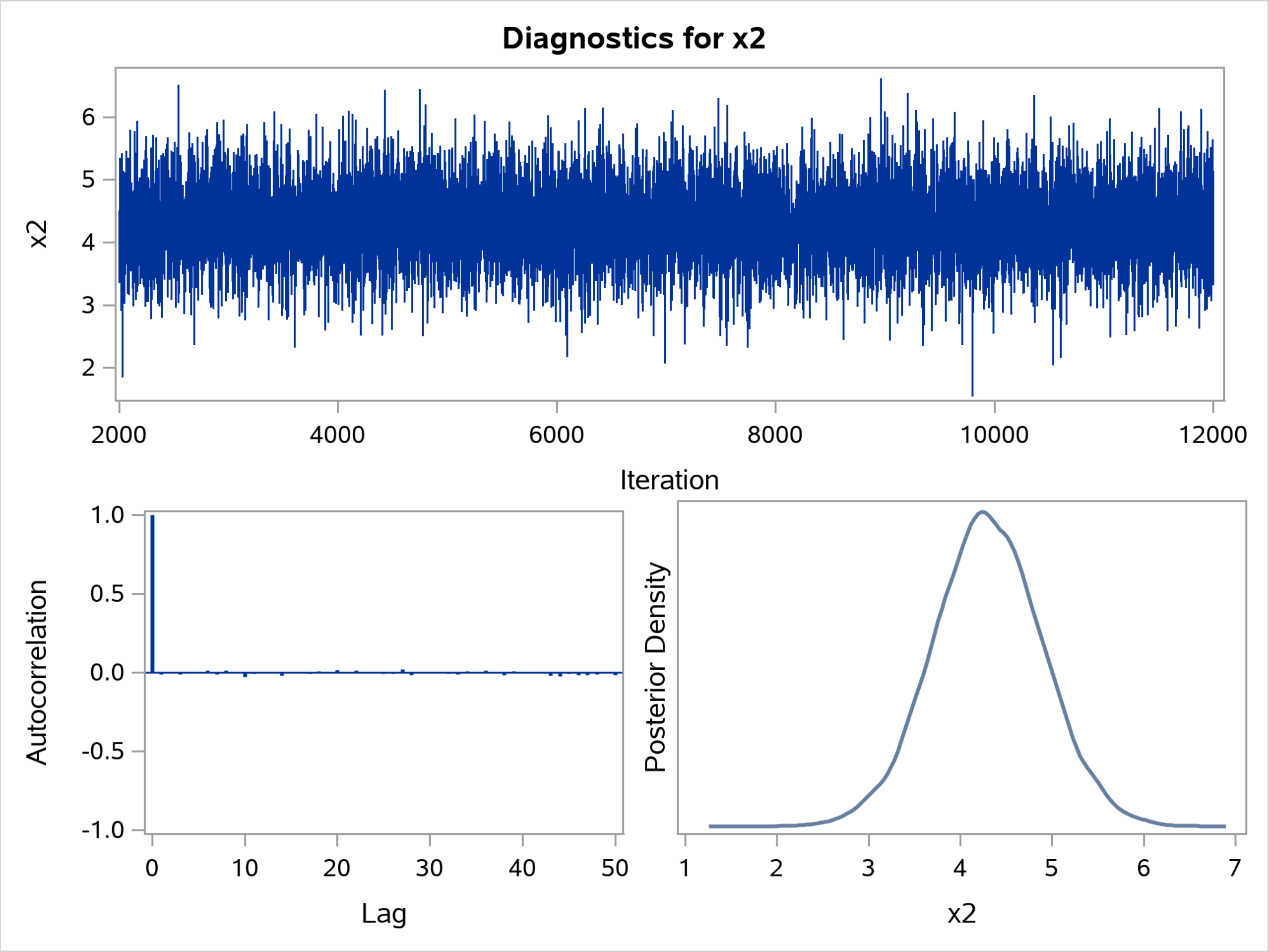


**Figure 51.23** Diagnostic Plots for X3

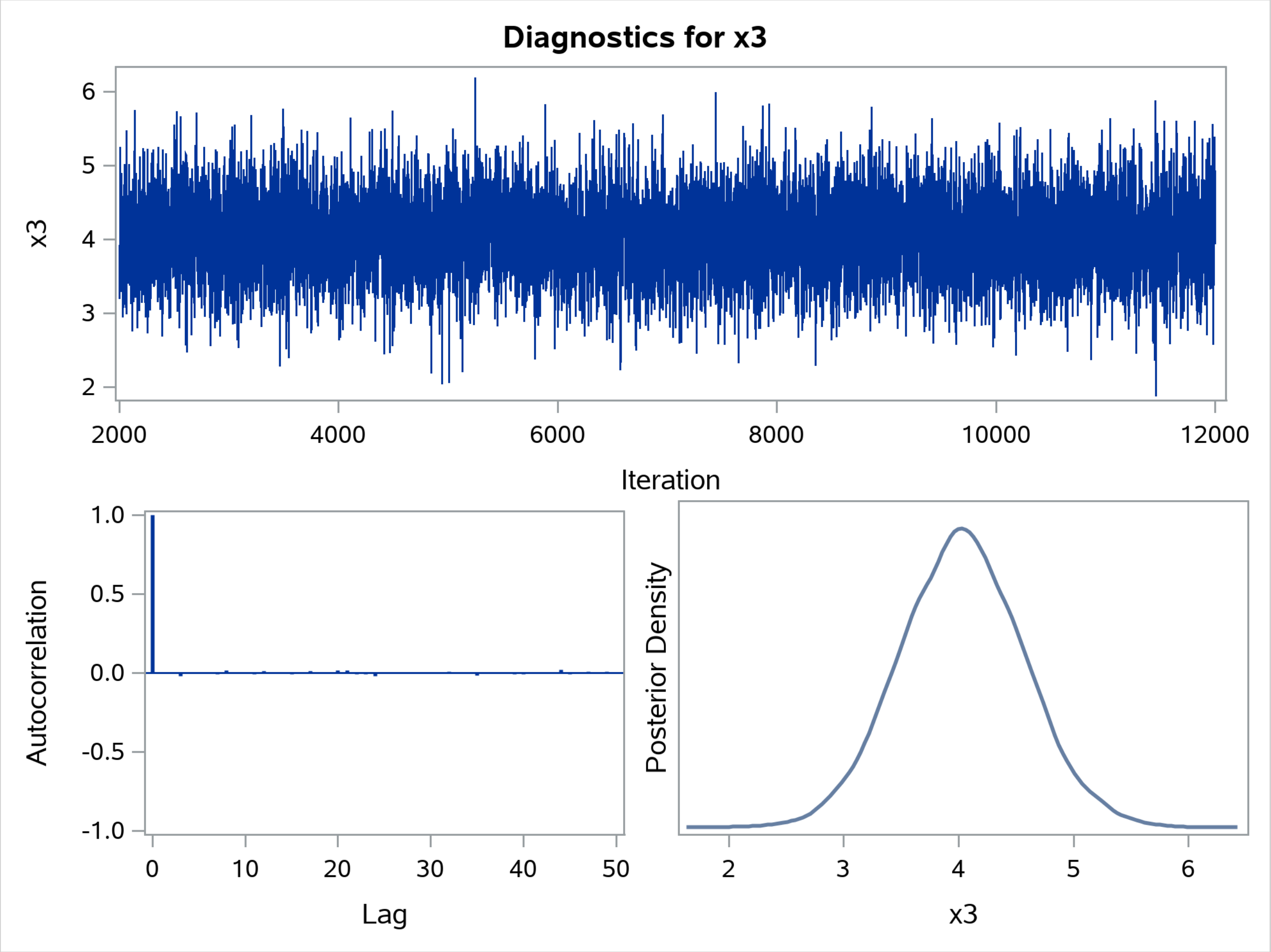


**Figure 51.24** Diagnostic Plots for X4

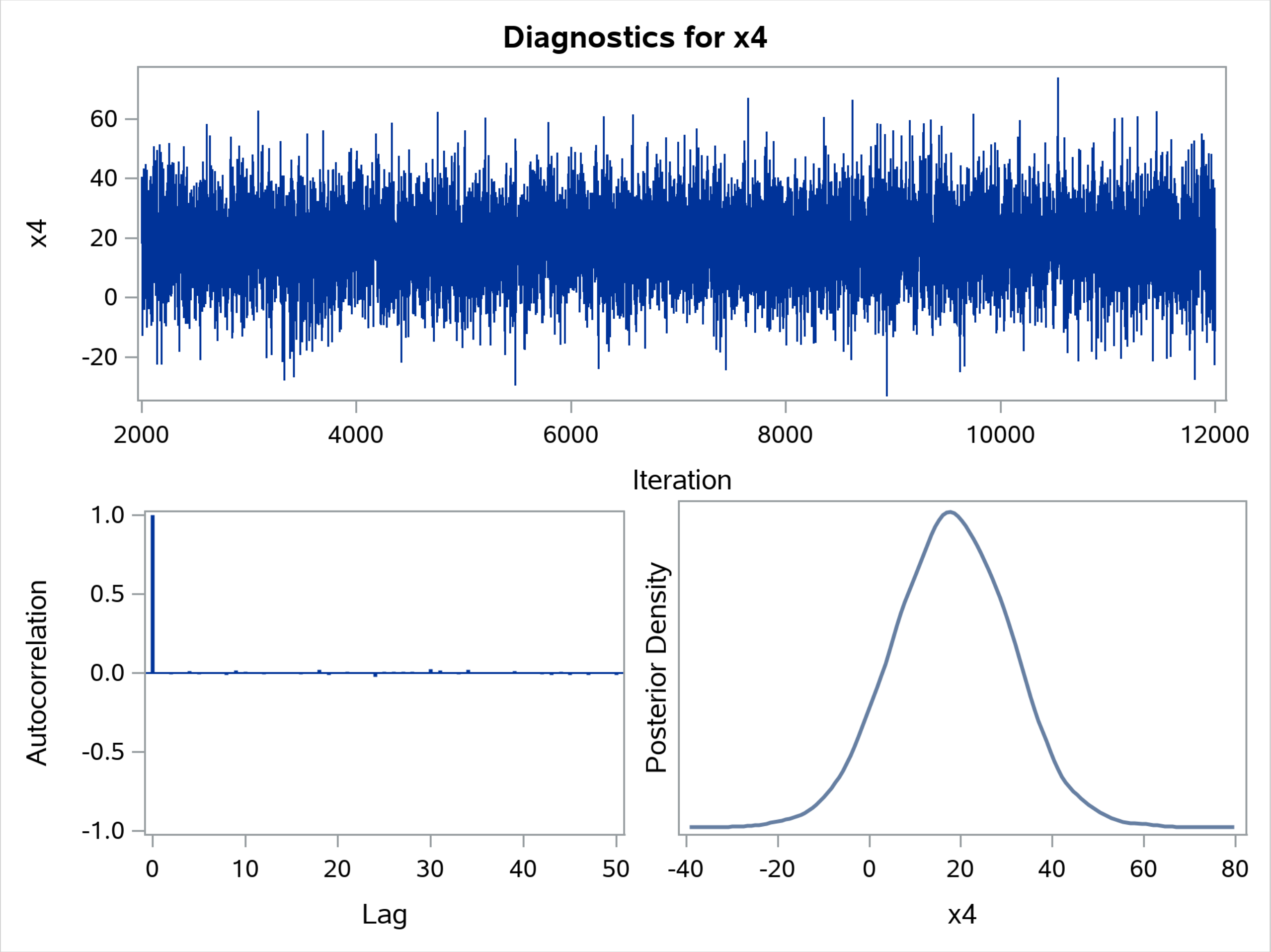
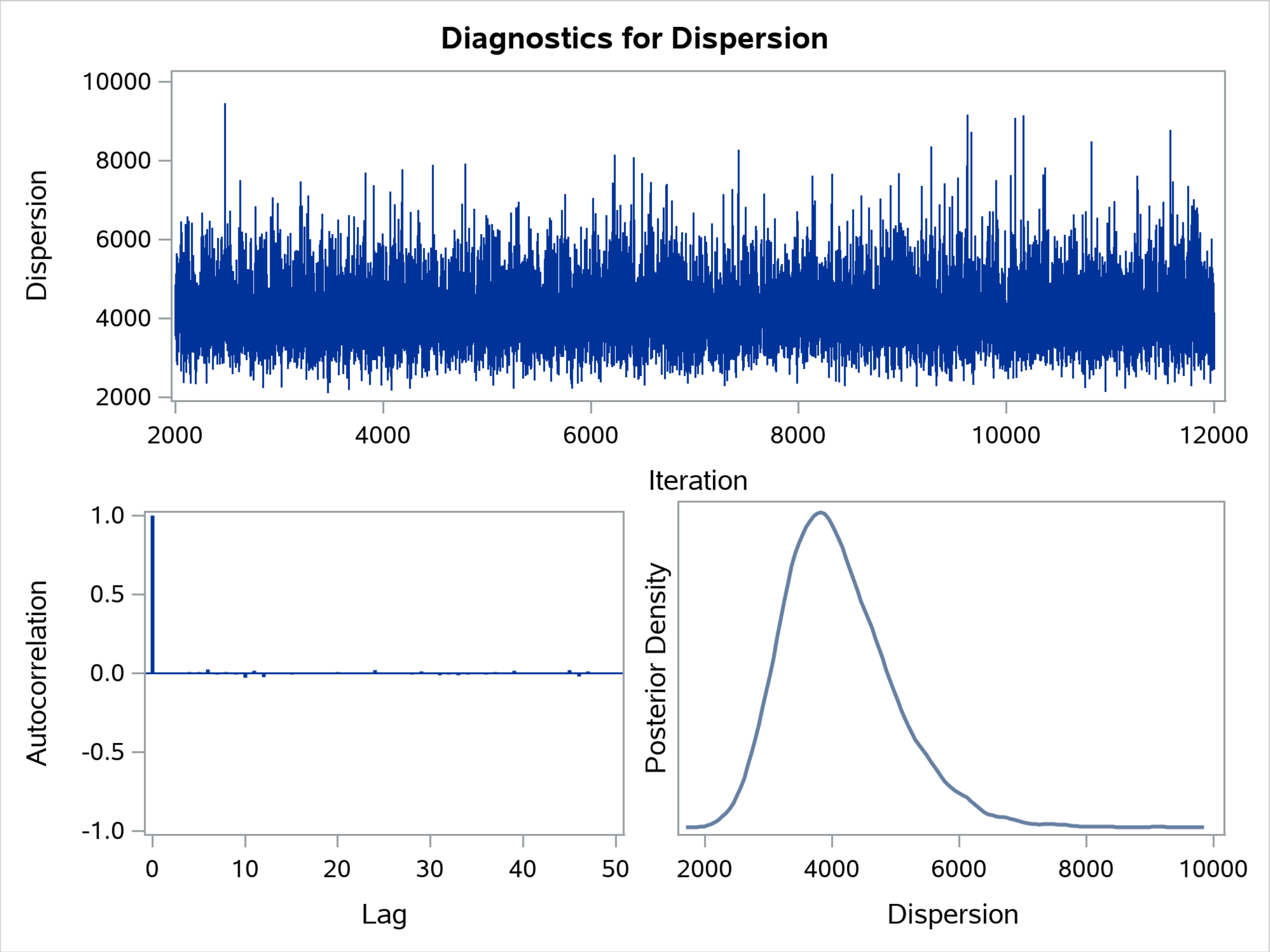


**Figure 51.25** Diagnostic Plots for X5



Suppose, for illustration, a question of scientific interest is whether blood clotting score has a positive effect on survival time. Since the model parameters are regarded as random quantities in a Bayesian analysis, you can answer this question by estimating the conditional probability of $\beta_1$ being positive, given the data, $\Pr(\beta_1 > 0|\mathbf{Y})$, from the posterior distribution samples. The following SAS statements compute the estimate of the probability of $\beta_1$ being positive:

```
data Prob;
   set PostSurg;
   Indicator = (logX1 > 0);
   label Indicator= 'log(Blood Clotting Score) > 0';
run;

proc Means data = Prob(keep=Indicator) n mean;
run;
```

As shown in Figure 51.26, there is a 1.00 probability of a positive relationship between the logarithm of a blood clotting score and survival time, adjusted for the other covariates.

**Figure 51.26** Probability That $\beta_1 > 0$

**The MEANS Procedure**

| Analysis Variable : Indicator log(Blood Clotting Score) > 0 | |
|---|---|
| **N** | **Mean** |
| 10000 | 0.9999000 |

## 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 et al. (1994). The binary response is the wheezing status of 16 children at ages 9, 10, 11, and 12 years. A value of 1 of wheezing status indicates the occurrence of wheezing. The mean response is modeled as a logistic regression model by 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(event='1') = 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 EVENT='1' option in the MODEL statement models the probability that wheeze = 1. If the EVENT='1' option had not been specified, the probability that wheeze = 0 would be modeled by default.

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 be 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 valid only for the initial model fit. The following figures display information that applies to the GEE model fit.

Figure 51.27 displays general information about the GEE model fit.

**Figure 51.27** GEE Model Information

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

**Figure 51.28** GEE Parameter Estimate Covariance Matrices

**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 |

**Figure 51.28** *continued*

**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 |

The exchangeable working correlation matrix specified by the CORRW option is displayed in Figure 51.29.

**Figure 51.29** GEE Working Correlation Matrix

**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 |

The parameter estimates table, displayed in Figure 51.30, 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 that displays model-based standard errors can be created by using the REPEATED statement option MODELSE.

**Figure 51.30** GEE Parameter Estimates Table

**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 |

## Syntax: GENMOD Procedure

The following statements are available in the GENMOD procedure. Items within the < > are optional.

```
PROC GENMOD < options > ;
    ASSESS | ASSESSMENT VAR=(effect)| LINK < / options > ;
    BAYES < options > ;
    BY variables ;
    CLASS variable < (options) > . . . < variable < (options) > > < / options > ;
    CODE < options > ;
    CONTRAST 'label' contrast-specification < / options > ;
    DEVIANCE variable = expression ;
    EFFECTPLOT < plot-type < (plot-definition-options) > > < / options > ;
    ESTIMATE 'label' effect values < , . . . effect values > < / options > ;
    EXACT < 'label' > < INTERCEPT > < effects > < / options > ;
    EXACTOPTIONS options ;
    FREQ | FREQUENCY variable ;
    FWDLINK variable = expression ;
    INVLINK variable = expression ;
    LSMEANS < model-effects > < / options > ;
    LSMESTIMATE model-effect < 'label' > values < divisor=n > < , . . . < 'label' > values < divisor=n > >
                    < / options > ;
    MODEL response = < effects  > < / options > ;
    OUTPUT < OUT=SAS-data-set > < keyword=name . . . keyword=name > ;
    Programming statements  ;
    REPEATED SUBJECT=subject-effect < / options > ;
    SLICE model-effect < / options > ;
    STORE < OUT= >item-store-name < / LABEL='label' > ;
    STRATA variable < (option) > . . . < variable < (option) > > < / options > ;
    WEIGHT | SCWGT variable ;
    VARIANCE variable = expression  ;
    ZEROMODEL < effects  > < / options > ;
```

The ASSESS, BAYES, BY, CLASS, CODE, CONTRAST, DEVIANCE, ESTIMATE, FREQUENCY, FWDLINK, INVLINK, MODEL, OUTPUT, programming statements, REPEATED, VARIANCE, WEIGHT, and ZEROMODEL statements are described in full after the PROC GENMOD statement in alphabetical order. The EFFECTPLOT, LSMEANS, LSMESTIMATE, SLICE, and STORE statements are common to many procedures. Summary descriptions of functionality and syntax for these statements are also given after the PROC GENMOD statement in alphabetical order, and full documentation about them is available in Chapter 20, "Shared Concepts and Topics."

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

## PROC GENMOD Statement

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

The PROC GENMOD statement invokes the GENMOD procedure. Table 51.1 summarizes the *options* available in the PROC GENMOD statement.

**Table 51.1** PROC GENMOD Statement Options

| Option | Description |
|---|---|
| DATA= | Specifies the input data set |
| DESCENDING | Sorts response variable in the reverse of the default order |
| EXACTONLY | Requests only the exact analyses |
| NAMELEN= | Specifies the length of effect names |
| ORDER= | Specifies the sort order of CLASS variable |
| PLOTS | Controls the plots produced through ODS Graphics |
| RORDER= | Specifies the sort order for the levels of the response variable |

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**

**DESCEND**

**DESC**

specifies that the levels of the response variable for the ordinal multinomial model and the binomial model with single variable response syntax 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.

**EXACTONLY**

requests only the exact analyses. The asymptotic analysis that PROC GENMOD usually performs is suppressed.

**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=DATA | FORMATTED | FREQ | INTERNAL**

specifies the sort order for the levels of the classification variables (which are specified in the CLASS statement).

The ORDER= option can be useful when you use the CONTRAST or ESTIMATE statement because it determines which parameters in the model correspond to each level in the data.

This option applies to the levels for all classification variables, except when you use the (default) ORDER=FORMATTED option with numeric classification variables that have no explicit format. In that case, the levels of such variables are ordered by their internal value.

The ORDER= option can take the following values:

| **Value of ORDER=** | **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 about sort order, see the chapter on the SORT procedure in the *Base SAS Procedures Guide* and the discussion of BY-group processing in the “Grouping Data” section of *SAS Programmers Guide: Essentials*.

**PLOTS** < (*global-plot-options*) > < =*plot-request* < (*options*) > >

**PLOTS** < (*global-plot-options*) > < =(*plot-request* < (*options*) > < ... *plot-request* < (*options*) > >) >

specifies plots to be created using ODS Graphics. Many of the observational statistics in the output data set can be plotted using this option. You are not required to create an output data set in order to produce a plot. When you specify only one plot request, you can omit the parentheses around the plot request. Here are some examples:

```
plots=all
plots=predicted
plots=(predicted reschi)
plots(unpack)=dfbeta
```

ODS Graphics must be enabled before plots can be requested. For example:

```
proc genmod plots=all;
   model y = x;
run;
```

For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” on page 687 in Chapter 24, “Statistical Graphics Using ODS.”

Any specified *global-plot-options* apply to all plots that are specified with *plot-requests*. The following *global-plot-options* are available.

**CLUSTERLABEL**

displays formatted levels of the SUBJECT= effect instead of plot symbols. This option applies only to diagnostic statistics for models fit by GEEs that are plotted against cluster number, and provides a way to identify cluster level names with corresponding ordered cluster numbers.

**UNPACK**

displays multiple plots individually. The default is to display related multiple plots in a panel.

See the section "OUTPUT Statement" on page 3627 for definitions of the statistics specified with the *plot-requests*. The *plot-requests* include the following:

**ALL**

produces all available plots.

**COOKSD**

**DOBS**

plots the Cook's distance statistic as a function of observation number.

**DFBETA**

plots the $\boldsymbol{\beta}$ deletion statistic as a function of observation number for each regression parameter in the model.

**DFBETAS**

plots the standardized $\boldsymbol{\beta}$ deletion statistic as a function of observation number for each regression parameter in the model.

**LEVERAGE**

plots the leverage as a function of observation number.

**OVERDISP**

plots the predicted variance as a function of the predicted mean for a zero-inflated response.

**PREDICTED< (*option*) >**

plots predicted values with confidence limits as a function of observation number. The PREDICTED plot request has the following *option*:

**CLM**

includes confidence limits in the predicted value plot.

**PZERO**

plots the zero inflation probability for zero-inflated Poisson and negative binomial models as a function of observation number.

**RESCHI< (*options*) >**

The RESCHI plot request has the following *options*:

**INDEX**

plots as a function of observation number.

**XBETA**
plots as a function of linear predictor.

If you do not specify an *option*, Pearson residuals are plotted as a function of observation number.

**RESDEV< (*options*) >**
plots deviance residuals. The RESDEV plot request has the following *options*:

**INDEX**
plots as a function of observation number.

**XBETA**
plots as a function of linear predictor.

If you do not specify an *option*, deviance residuals are plotted as a function of observation number.

**RESLIK< (*options*) >**
plots likelihood residuals. The RESLIK plot request has the following *options*:

**INDEX**
plots as a function of observation number.

**XBETA**
plots as a function of linear predictor.

If you do not specify an *option*, likelihood residuals are plotted as a function of observation number.

**RESRAW< (*options*) >**
plots raw residuals. The RESRAW plot request has the following *options*:

**INDEX**
plots as a function of observation number.

**XBETA**
plots as a function of linear predictor.

If you do not specify an *option*, raw residuals are plotted as a function of observation number.

**STDRESCHI< (*options*) >**
plots standardized Pearson residuals. The STDRESCHI plot request has the following *options*:

**INDEX**
plots as a function of observation number.

**XBETA**
plots as a function of linear predictor.

If you do not specify an *option*, standardized Pearson residuals are plotted as a function of observation number.

**STDRESDEV< (*options*) >**

plots standardized deviance residuals. The STDRESDEV plot request has the following *options*:

**INDEX**

plots as a function of observation number.

**XBETA**

plots as a function of linear predictor.

If you do not specify an *option*, standardized deviance residuals are plotted as a function of observation number.

If you fit a model by using generalized estimating equations (GEEs), the following additional *plot-requests* are available:

**CLEVERAGE**

plots the cluster leverage as a function of ordered cluster.

**CLUSTERCOOKSD**

**DCLS**

plots the cluster Cook's distance statistic as a function of ordered cluster.

**CLUSTERDFIT**

**MCLS**

plots the studentized cluster Cook's distance statistic as a function of ordered cluster.

**DFBETAC**

plots the cluster deletion statistic as a function of ordered cluster for each regression parameter in the model.

**DFBETACS**

plots the standardized cluster deletion statistic as a function of ordered cluster for each regression parameter in the model.

**RORDER=*keyword***

specifies the sort order for the levels of the response variable. This order determines which intercept parameter in the model corresponds to each level in the data. If RORDER=FORMATTED for numeric variables for which you have supplied no explicit format, the levels are ordered by their internal values. 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 about sort order, see the chapter on the SORT procedure in the *Base 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 turn displayed output back on with the statement `ODS SELECT ALL;`. See Table 51.13 and Table 51.14 for the names of output tables available from PROC GENMOD. For more information about ODS, see Chapter 23, "Using the Output Delivery System."

## ASSESS Statement

**ASSESS VAR=(***effect***)| LINK** < / *options* > **;**

**ASSESSMENT VAR=(***effect***)| LINK** < / *options* > **;**

The ASSESS statement computes and plots, using ODS Graphics, model-checking statistics based on aggregates of residuals. See the section "Assessment of Models Based on Aggregates of Residuals" on page 3672 for details about the model assessment methods available in GENMOD.

The types of aggregates available are cumulative residuals, moving sums of residuals, and loess smoothed residuals. If you do not specify which aggregate to use, the assessments are based on cumulative sums. PROC GENMOD uses ODS Graphics for graphical displays. For specific information about the graphics available in PROC GENMOD, see the section "ODS Graphics" on page 3703.

You must specify either LINK or VAR= in order to create an analysis.

**LINK** requests the assessment of the link function by performing the analysis with respect to the linear predictor.

**VAR=(***effect***)** specifies that the functional form of a covariate be checked by performing the analysis with respect to the variable identified by the effect. The effect must be specified in the MODEL statement and must contain only continuous variables (variables not listed in a CLASS statement).

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

**CRPANEL**

requests that a plot with four panels showing just a few of the paths from the default aggregate plot to make it easier to compare simulated and observed paths. The plot in each panel contains aggregates of the observed residuals and two simulated curves (fewer if NPATHS= is less than 8).

**LOESS**< **(***number***)** >

**LOWESS**< **(***number***)** >

requests model assessment based on loess smoothed residuals with optional *number* the fraction of data used; *number* must be between zero and one. If *number* is not specified, the default value one-third is used.

**NPATHS=***number*

**NPATH=***number*

**PATHS=***number*

**PATH=***number*

specifies the number of simulated paths to plot in the default aggregate residuals plot. The default value of *number* is twenty.

**RESAMPLE< =***number* **>**

**RESAMPLES< =***number* **>**

specifies that a $p$-value be computed based on 1,000 simulated paths, or *number* paths, if *number* is specified.

**SEED=***number*

specifies a seed for the normal random number generator used in creating simulated realizations of aggregates of residuals for plots and estimating $p$-values. Specifying a seed enables you to produce identical graphs and $p$-values from one run of the procedure to the next run. If a seed is not specified, or if *number* is negative or zero, a random number seed is derived from the time of day.

**WINDOW< (***number***) >**

requests assessment based on a moving sum window of width *number*. If *number* is not specified, a value of one-half of the range of the $x$-coordinate is used.

## BAYES Statement

**BAYES** < *options* > **;**

The BAYES statement requests a Bayesian analysis of the regression model by using Gibbs sampling. The Bayesian posterior samples (also known as the chain) for the regression parameters are not tabulated. The Bayesian posterior samples (also known as the chain) for the regression parameters can be output to a SAS data set. Table 51.2 summarizes the *options* available in the BAYES statement.

**Table 51.2** BAYES Statement Options

| Option | Description |
|---|---|
| **Monte Carlo Options** | |
| INITIAL= | Specifies the initial values of the chain |
| INITIALMLE | Specifies that maximum likelihood estimates be used as initial values of the chain |
| METROPOLIS= | Specifies the use of a Metropolis step in the ARMS algorithm |
| NBI= | Specifies the number of burn-in iterations |
| NMC= | Specifies the number of iterations after burn-in |
| SAMPLING= | Specifies the algorithm used to sample the posterior distribution |
| SEED= | Specifies the random number generator seed |
| THINNING= | Controls the thinning of the Markov chain |

**Table 51.2** *continued*

| Option | Description |
|---|---|
| **Model and Prior Options** | |
| COEFFPRIOR= | Specifies the prior of the regression coefficients |
| DISPERSIONPRIOR= | Specifies the prior of the dispersion parameter |
| PRECISIONPRIOR= | Specifies the prior of the precision parameter |
| SCALEPRIOR= | Specifies the prior of the scale parameter |
| **Summary Statistics and Convergence Diagnostics** | |
| DIAGNOSTICS= | Displays convergence diagnostics |
| PLOTS= | Displays diagnostic plots |
| STATISTICS= | Displays summary statistics of the posterior samples |
| **Posterior Samples** | |
| OUTPOST= | Names a SAS data set for the posterior samples |

The following list describes these *options* and their *suboptions*.

**COEFFPRIOR=JEFFREYS< (*option*) > | NORMAL< (*options*) > | UNIFORM**

**COEFF=JEFFREYS< (*options*) > | NORMAL< (*options*) > | UNIFORM**

**CPRIOR=JEFFREYS< (*options*) > | NORMAL< (*options*) > | UNIFORM**

specifies the prior distribution for the regression coefficients. The default is COEFFPRIOR=UNIFORM, which specifies the noninformative and improper prior of a constant.

Jeffreys' prior is specified by COEFFPRIOR=JEFFREYS, which can be followed by the following option in parentheses. Jeffreys' prior is proportional to $|I(\boldsymbol{\beta})|^{\frac{1}{2}}$, where $I(\boldsymbol{\beta})$ is the Fisher information matrix. See the section "Jeffreys' Prior" on page 3681 and Ibrahim and Laud (1991) for more details.

**CONDITIONAL**

specifies that the Jeffreys' prior, conditional on the current Markov chain value of the generalized linear model precision parameter $\tau$, is proportional to $|\tau\mathbf{I}(\boldsymbol{\beta})|^{\frac{1}{2}}$.

The normal prior is specified by COEFFPRIOR=NORMAL, which can be followed by one of the following *options* enclosed in parentheses. However, if you do not specify an *option*, the normal prior $N(\mathbf{0}, 10^6\mathbf{I})$, where $\mathbf{I}$ is the identity matrix, is used. See the section "Normal Prior" on page 3682 for more details.

**CONDITIONAL**

specifies that the normal prior, conditional on the current Markov chain value of the generalized linear model precision parameter $\tau$, is $N(\boldsymbol{\mu}, \tau^{-1}\boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance of the normal prior specified by other normal options.

**INPUT=***SAS-data-set*

specifies a SAS data set containing the mean and covariance information of the normal prior. The data set must have a _TYPE_ variable to represent the type of each observation and a variable for each regression coefficient. If the data set also contains a _NAME_ variable, the values of this variable are used to identify the covariances for the _TYPE_='COV' observations; otherwise, the _TYPE_='COV' observations are assumed to be in the same order as the explanatory variables in the MODEL statement. PROC GENMOD reads the mean vector from the observation with _TYPE_='MEAN' and reads the covariance matrix from observations with _TYPE_='COV'.

For an independent normal prior, specify either variances or a covariance matrix in addition to the means. You can specify the variances in the _TYPE_='VAR' option, or you can equivalently specify the precisions (inverse of the variances) in the _TYPE_='PRECISION' option. Each variable in the INPUT= data set corresponds to a parameter in the model, and the variable name must match the parameter name. Parameter names can be found in any of the tables (such as the "Initial Values of the Chain" table) in the Bayesian Analysis section of the results. See Example 51.10 for an example.

**RELVAR< =*c* >**

specifies the normal prior $N(\mathbf{0}, c\mathbf{J})$, where $\mathbf{J}$ is a diagonal matrix with diagonal elements equal to the variances of the corresponding ML estimator. By default, $c = 10^6$.

**VAR< =*c* >**

specifies the normal prior $N(\mathbf{0}, c\mathbf{I})$, where $\mathbf{I}$ is the identity matrix.

**DIAGNOSTICS=ALL | NONE | (*keyword-list*)**

**DIAG=ALL | NONE | (*keyword-list*)**

controls the number of diagnostics produced. You can request all the following diagnostics by specifying DIAGNOSTICS=ALL. If you do not want any of these diagnostics, specify DIAGNOSTICS=NONE. If you want some but not all of the diagnostics, or if you want to change certain settings of these diagnostics, specify a subset of the following *keywords*. The default is DIAGNOSTICS=(AUTOCORR ESS GEWEKE).

**AUTOCORR < (LAGS= *numeric-list*) >**

computes the autocorrelations of lags given by LAGS= list for each parameter. Elements in the list are truncated to integers and repeated values are removed. If the LAGS= option is not specified, autocorrelations of lags 1, 5, 10, and 50 are computed for each variable. See the section "Autocorrelations" on page 174 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for details.

**ESS**

computes Carlin's estimate of the effective sample size, the correlation time, and the efficiency of the chain for each parameter. See the section "Effective Sample Size" on page 175 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for details.

**GELMAN < (*gelman-options*) >**

computes the Gelman and Rubin convergence diagnostics. You can specify one or more of the following *gelman-options*:

**NCHAIN | N=*number***

specifies the number of parallel chains used to compute the diagnostic, and must be 2 or larger. The default is NCHAIN=3. If an INITIAL= data set is used, NCHAIN defaults to the number of rows in the INITIAL= data set. If any number other than this is specified with the NCHAIN= option, the NCHAIN= value is ignored.

**ALPHA=***value*

specifies the significance level for the upper bound. The default is ALPHA=0.05, resulting in a 97.5% bound.

See the section "Gelman and Rubin Diagnostics" on page 167 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for details.

**GEWEKE < (***geweke-options***) >**

computes the Geweke spectral density diagnostics, which are essentially a two-sample $t$ test between the first $f_1$ portion and the last $f_2$ portion of the chain. The default is $f_1 = 0.1$ and $f_2 = 0.5$, but you can choose other fractions by using the following *geweke-options*:

**FRAC1=***value*

specifies the fraction $f_1$ for the first window.

**FRAC2=***value*

specifies the fraction $f_2$ for the second window.

See the section "Geweke Diagnostics" on page 169 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for details.

**HEIDELBERGER < (***heidel-options***) >**

computes the Heidelberger and Welch diagnostic for each variable, which consists of a stationarity test of the null hypothesis that the sample values form a stationary process. If the stationarity test is not rejected, a halfwidth test is then carried out. Optionally, you can specify one or more of the following *heidel-options*:

**SALPHA=***value*

specifies the $\alpha$ level $(0 < \alpha < 1)$ for the stationarity test.

**HALPHA=***value*

specifies the $\alpha$ level $(0 < \alpha < 1)$ for the halfwidth test.

**EPS=***value*

specifies a positive number $\epsilon$ such that if the halfwidth is less than $\epsilon$ times the sample mean of the retained iterates, the halfwidth test is passed.

See the section "Heidelberger and Welch Diagnostics" on page 170 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for details.

**MCSE**

**MCERROR**

computes the Monte Carlo standard error for each parameter. The Monte Caro standard error, which measures the simulation accuracy, is the standard error of the posterior mean estimate and is calculated as the posterior standard deviation divided by the square root of the effective sample size. See the section "Standard Error of the Mean Estimate" on page 176 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for details.

**RAFTERY**< **(***raftery-options***)** >

computes the Raftery and Lewis diagnostics that evaluate the accuracy of the estimated quantile ($\hat{\theta}_Q$ for a given $Q \in (0,1)$) of a chain. $\hat{\theta}_Q$ can achieve any degree of accuracy when the chain is allowed to run for a long time. A stopping criterion is when the estimated probability $\hat{P}_Q = \Pr(\theta \leq \hat{\theta}_Q)$ reaches within $\pm R$ of the value $Q$ with probability $S$; that is, $\Pr(Q - R \leq \hat{P}_Q \leq Q + R) = S$. The following *raftery-options* enable you to specify $Q, R, S$, and a precision level $\epsilon$ for the test:

**QUANTILE | Q=***value*

specifies the order (a value between 0 and 1) of the quantile of interest. The default is 0.025.

**ACCURACY | R=***value*

specifies a small positive number as the margin of error for measuring the accuracy of estimation of the quantile. The default is 0.005.

**PROBABILITY | S=***value*

specifies the probability of attaining the accuracy of the estimation of the quantile. The default is 0.95.

**EPSILON | EPS=***value*

specifies the tolerance level (a small positive number) for the stationary test. The default is 0.001.

See the section "Raftery and Lewis Diagnostics" on page 172 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for details.

**DISPERSIONPRIOR=GAMMA**< **(***options***)** > **| IGAMMA**< **(***options***)** > **| IMPROPER**

**DPRIOR=GAMMA**< **(***options***)** > **| IGAMMA**< **(***options***)** > **| IMPROPER**

specifies that Gibbs sampling be performed on the generalized linear model dispersion parameter and the prior distribution for the dispersion parameter, if there is a dispersion parameter in the model. For models that do not have a dispersion parameter (the Poisson and binomial), this option is ignored. Note that you can specify Gibbs sampling on either the dispersion parameter $\phi$, the scale parameter $\sigma = \phi^{\frac{1}{2}}$, or the precision parameter $\tau = \phi^{-1}$, with the DPRIOR=, SPRIOR=, and PPRIOR= options, respectively. These three parameters are transformations of one another, and you should specify Gibbs sampling for only one of them.

A gamma prior $G(a,b)$ with density $f(t) = \frac{b(bt)^{a-1}\mathrm{e}^{-bt}}{\Gamma(a)}$ is specified by DISPERSIONPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters $a$ and $b$ are the shape and inverse-scale parameters of the gamma distribution, respectively. See the section "Gamma Prior" on page 3681 for details. The default is $G(10^{-4}, 10^{-4})$.

**RELSHAPE**< **=***c* >

specifies independent $G(c\hat{\phi}, c)$ distribution, where $\hat{\phi}$ is the MLE of the dispersion parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\phi}$ and the variance is $\frac{\hat{\phi}}{c}$. By default, *c*=$10^{-4}$.

**SHAPE=*a***

**ISCALE=*b***

when both specified, results in a $G(a,b)$ prior.

**SHAPE=*c***

when specified alone, results in a $G(c,c)$ prior.

**ISCALE=*c***

when specified alone, results in a $G(c,c)$ prior.

An inverse gamma prior $\mathrm{IG}(a,b)$ with density $f(t) = \frac{b^a}{\Gamma(a)} t^{-(a+1)} \mathrm{e}^{-b/t}$ is specified by DISPERSIONPRIOR=IGAMMA, which can be followed by one of the following inverse gamma *options* enclosed in parentheses. The hyperparameters $a$ and $b$ are the shape and scale parameters of the inverse gamma distribution, respectively. See the section "Inverse Gamma Prior" on page 3681 for details. The default is $\mathrm{IG}(2.001, 0.001)$.

**RELSHAPE< =*c*>**

specifies independent $\mathrm{IG}(\frac{c+\hat{\phi}}{\hat{\phi}}, c)$ distribution, where $\hat{\phi}$ is the MLE of the dispersion parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\phi}$. By default, *c*=$10^{-4}$.

**SHAPE=*a***

**SCALE=*b***

when both specified, results in a $\mathrm{IG}(a,b)$ prior.

**SHAPE=*c***

when specified alone, results in an $\mathrm{IG}(c,c)$ prior.

**SCALE=*c***

when specified alone, results in an $\mathrm{IG}(c,c)$ prior.

An improper prior with density $f(t)$ proportional to $t^{-1}$ is specified with DISPERSIONPRIOR=IMPROPER.

**INITIAL=*SAS-data-set***

specifies the SAS data set that contains the initial values of the Markov chains. The INITIAL= data set must contain all the variables of the model. You can specify multiple rows as the initial values of the parallel chains for the Gelman-Rubin statistics, but posterior summaries, diagnostics, and plots are computed only for the first chain. If the data set also contains the variable _SEED_, the value of the _SEED_ variable is used as the seed of the random number generator for the corresponding chain.

**INITIALMLE**

specifies that maximum likelihood estimates of the model parameters be used as initial values of the Markov chain. If this option is not specified, estimates of the mode of the posterior distribution obtained by optimization are used as initial values.

**METROPOLIS=YES | NO**

specifies the use of a Metropolis step to generate Gibbs samples for posterior distributions that are not log concave. The default value is METROPOLIS=YES.

**NBI=***number*

specifies the number of burn-in iterations before the chains are saved. The default is 2000.

**NMC=***number*

specifies the number of iterations after the burn-in. The default is 10000.

**OUTPOST=***SAS-data-set*

**OUT=***SAS-data-set*

names the SAS data set that contains the posterior samples. See the sections "OUTPOST= Output Data Set" on page 3683 and "Posterior Samples Output Data Set" on page 3680 for more information. Alternatively, you can create the output data set by specifying an ODS OUTPUT statement as follows:

**ODS OUTPUT POSTERIORSAMPLE=***SAS-data-set*

**PRECISIONPRIOR=GAMMA< (***options***) > | IMPROPER**

**PPRIOR=GAMMA< (***options***) > | IMPROPER**

specifies that Gibbs sampling be performed on the generalized linear model precision parameter and the prior distribution for the precision parameter, if there is a precision parameter in the model. For models that do not have a precision parameter (the Poisson and binomial), this option is ignored. Note that you can specify Gibbs sampling on either the dispersion parameter $\phi$, the scale parameter $\sigma = \phi^{\frac{1}{2}}$, or the precision parameter $\tau = \phi^{-1}$, with the DPRIOR=, SPRIOR=, and PPRIOR= options, respectively. These three parameters are transformations of one another, and you should specify Gibbs sampling for only one of them.

A gamma prior $G(a,b)$ with density $f(t) = \frac{b(bt)^{a-1}\mathrm{e}^{-bt}}{\Gamma(a)}$ is specified by PRECISIONPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters $a$ and $b$ are the shape and inverse-scale parameters of the gamma distribution, respectively. See the section "Gamma Prior" on page 3681 for details. The default is $G(10^{-4}, 10^{-4})$.

**RELSHAPE< =***c***>**

specifies independent $G(c\hat{\tau}, c)$ distribution, where $\hat{\tau}$ is the MLE of the dispersion parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\tau}$ and the variance is $\frac{\hat{\tau}}{c}$. By default, $c = 10^{-4}$.

**SHAPE=***a*

**ISCALE=***b*

when both specified, results in a $G(a,b)$ prior.

**SHAPE=***c*

when specified alone, results in an $G(c,c)$ prior.

**ISCALE=***c*

when specified alone, results in an $G(c,c)$ prior.

An improper prior with density $f(t)$ proportional to $t^{-1}$ is specified with PRECISIONPRIOR=IMPROPER.

**PLOTS< (***global-plot-options***) >=***plot-request*

**PLOTS< (***global-plot-options***) >=(***plot-request* **< . . .** *plot-request* **>)**

controls the display of diagnostic plots. Three types of plots can be requested: trace plots, autocorrelation function plots, and kernel density plots. By default, the plots are displayed in panels unless the *global-plot-option* UNPACK is specified. Also, when you are specifying more than one type of plots, the plots are displayed by parameters unless the *global-plot-option* GROUPBY is specified. When you specify only one *plot-request*, you can omit the parentheses around the *plot-request*. For example:

```
plots=none
plots(unpack)=trace
plots=(trace autocorr)
```

ODS Graphics must be enabled before requesting plots. For example, the following SAS statements enable ODS Graphics:

```
ods graphics on;
proc genmod;
   model y=x;
   bayes plots=trace;
run;
ods graphics off;
```

The *global-plot-options* are as follows:

**FRINGE**

creates a fringe plot on the X axis of the density plot.

**GROUPBY=PARAMETER**

**GROUPBY=TYPE**

specifies how the plots are grouped when there is more than one type of plot.

**GROUPBY=TYPE**

specifies that the plots be grouped by type.

**GROUPBY=PARAMETER**

specifies that the plots be grouped by parameter.

GROUPBY=PARAMETER is the default.

**LAGS=***n*

specifies that autocorrelations be plotted up to lag *n*. If this option is not specified, autocorrelations are plotted up to lag 50.

**SMOOTH**

displays a fitted penalized B-spline curve for each trace plot.

**UNPACKPANEL**

**UNPACK**

specifies that all paneled plots be unpacked, meaning that each plot in a panel is displayed separately.

The *plot-requests* include the following:

**ALL**

specifies all types of plots. PLOTS=ALL is equivalent to specifying PLOTS=(TRACE AUTOCORR DENSITY).

**AUTOCORR**

displays the autocorrelation function plots for the parameters.

**DENSITY**

displays the kernel density plots for the parameters.

**NONE**

suppresses all diagnostic plots.

**TRACE**

displays the trace plots for the parameters. See the section "Visual Analysis via Trace Plots" on page 163 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for details.

**SAMPLING=*option***

specifies an algorithm used to sample the posterior distribution. The following *options* are available:

**ARMS**

**GIBBS**

use the ARMS algorithm.

**GAMERMAN**

**GAM**

use the Gamerman algorithm. This is the default method.

**IM**

use the independent Metropolis algorithm.

**SCALEPRIOR=GAMMA< (*options*) > | IMPROPER**

**SPRIOR=GAMMA< (*options*) > | IMPROPER**

specifies that Gibbs sampling be performed on the generalized linear model scale parameter and the prior distribution for the scale parameter, if there is a scale parameter in the model. For models that do not have a scale parameter (the Poisson and binomial), this option is ignored. Note that you can specify Gibbs sampling on either the dispersion parameter $\phi$, the scale parameter $\sigma = \phi^{\frac{1}{2}}$, or the precision parameter $\tau = \phi^{-1}$, with the DPRIOR=, SPRIOR=, and PPRIOR= options, respectively. These three parameters are transformations of one another, and you should specify Gibbs sampling for only one of them.

A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}\mathrm{e}^{-bt}}{\Gamma(a)}$ is specified by SCALEPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters $a$ and $b$ are the shape and inverse-scale parameters of the gamma distribution, respectively. See the section "Gamma Prior" on page 3681 for details. The default is $G(10^{-4}, 10^{-4})$.

**RELSHAPE< =*c* >**

specifies independent $G(c\hat{\sigma}, c)$ distribution, where $\hat{\sigma}$ is the MLE of the dispersion parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\sigma}$ and the variance is $\frac{\hat{\sigma}}{c}$. By default, $c = 10^{-4}$.

**SHAPE=*a***

**ISCALE=*b***

when both specified, results in a $G(a, b)$ prior.

**SHAPE=*c***

when specified alone, results in an $G(c, c)$ prior.

**ISCALE=*c***

when specified alone, results in an $G(c, c)$ prior.

An improper prior with density $f(t)$ proportional to $t^{-1}$ is specified with SCALEPRIOR=IMPROPER.

**SEED=*number***

specifies an integer seed in the range 1 to $2^{31} - 1$ for the random number generator in the simulation. Specifying a seed enables you to reproduce identical Markov chains for the same specification. If the SEED= option is not specified, or if you specify a nonpositive seed, a random seed is derived from the time of day.

**STATISTICS < (*global-options*) > = ALL | NONE | *keyword* | (*keyword-list*)**

**STATS < (*global-options*) > = ALL | NONE | *keyword* | (*keyword-list*)**

controls the number of posterior statistics produced. Specifying STATISTICS=ALL is equivalent to specifying STATISTICS= (SUMMARY INTERVAL COV CORR). If you do not want any posterior statistics, you specify STATISTICS=NONE. The default is STATISTICS=(SUMMARY INTERVAL). See the section "Summary Statistics" on page 175 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for details. The *global-options* include the following:

**ALPHA=*numeric-list***

controls the probabilities of the credible intervals. The ALPHA= values must be between 0 and 1. Each ALPHA= value produces a pair of 100(1–ALPHA)% equal-tail and HPD intervals for each parameters. The default is the value of the ALPHA= option in the MODEL statement, or 0.05 if that option is not specified (yielding the 95% credible intervals for each parameter).

**PERCENT=*numeric-list***

requests the percentile points of the posterior samples. The PERCENT= values must be between 0 and 100. The default is PERCENT=25, 50, 75, which yield the 25th, 50th, and 75th percentile points, respectively, for each parameter.

The list of *keywords* includes the following:

**CORR**

produces the posterior correlation matrix.

**COV**

produces the posterior covariance matrix.

**SUMMARY**

produces the means, standard deviations, and percentile points for the posterior samples. The default is to produce the 25th, 50th, and 75th percentile points, but you can use the global PERCENT= option to request specific percentile points.

**INTERVAL**

produces equal-tail credible intervals and HPD intervals. The default is to produce the 95% equal-tail credible intervals and 95% HPD intervals, but you can use the global ALPHA= option to request intervals of any probabilities.

**THINNING=***number*

**THIN=***number*

controls the thinning of the Markov chain. Only one in every $k$ samples is used when THINNING=$k$, and if NBI=$n_0$ and NMC=$n$, the number of samples kept is

$$\left[\frac{n_0+n}{k}\right]-\left[\frac{n_0}{k}\right]$$

where $[a]$ represents the integer part of the number $a$. The default is THINNING=1.

## BY Statement

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

You can specify a BY statement in PROC GENMOD to obtain separate analyses of observations in groups that are 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. If you specify more than one BY statement, only the last one specified is used.

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

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement in 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 by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the “Grouping Data” section of *SAS Programmers Guide: Essentials*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

## CLASS Statement

**CLASS** *variable* < **(***options***)** > . . . < *variable* < **(***options***)** > > < / *global-options* > **;**

The CLASS statement names the classification variables to be used as explanatory variables in the analysis. Response variables do not need to be specified in the CLASS statement.

The CLASS statement must precede the MODEL statement. Most options can be specified either as individual variable *options* or as *global-options*. You can specify *options* for each variable by enclosing the options in parentheses after the variable name. You can also specify *global-options* for the CLASS statement by placing them after a slash (/). *Global-options* are applied to all the variables that are specified in the CLASS statement. If you specify more than one CLASS statement, the *global-options* that are specified in any one CLASS statement apply to all CLASS statements. However, individual CLASS variable *options* override the *global-options*. You can specify the following values for either an *option* or a *global-option*:

**CPREFIX=***n*

specifies that, at most, the first *n* characters of a CLASS variable name be used in creating names for the corresponding design variables. The default is $32 - \min(32, \max(2, f))$, where $f$ is the formatted length of the CLASS variable.

**DESCENDING**

**DESC**

reverses the sort order of the classification variable. If you specify both the DESCENDING and ORDER= options, PROC GENMOD orders the categories according to the ORDER= option and then reverses that order.

**LPREFIX=***n*

specifies that, at most, the first *n* characters of a CLASS variable label be used in creating labels for the corresponding design variables. The default is $256 - \min(256, \max(2, f))$, where $f$ is the formatted length of the CLASS variable.

**MISSING**

treats missing values (., ._, .A, . . . , .Z for numeric variables and blanks for character variables) as valid values of the CLASS variable.

**ORDER=DATA | FORMATTED | FREQ | INTERNAL**

specifies the sort order for the levels of classification variables. This ordering determines which parameters in the model correspond to each level in the data, so this option can be useful when you use the CONTRAST statement. By default, ORDER=FORMATTED. For ORDER=FORMATTED and ORDER=INTERNAL, the sort order is machine-dependent. When ORDER=FORMATTED is in effect for numeric variables for which you have supplied no explicit format, the levels are ordered by their internal values.

The following table shows how PROC GENMOD interprets values of the ORDER= option:

| Value of ORDER= | Levels Sorted By |
|---|---|
| **DATA** | Order of appearance in the input data set |
| **FORMATTED** | External formatted values, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) values |
| **FREQ** | Descending frequency count; levels with more observations come earlier in the order |
| **INTERNAL** | Unformatted value |

For more information about sort order, see the chapter on the SORT procedure in the *Base SAS Procedures Guide* and the discussion of BY-group processing in the "Grouping Data" section of *SAS Programmers Guide: Essentials*.

**PARAM=***keyword*

specifies the parameterization method for the classification variable or variables. You can specify any of the *keywords* shown in the following table. The default is PARAM=GLM. Design matrix columns are created from CLASS variables according to the corresponding coding schemes.

| Value of PARAM= | Coding |
|---|---|
| **EFFECT** | Effect coding |
| **GLM** | Less-than-full-rank reference cell coding (this *keyword* can be used only in a global option) |
| **ORDINAL**<br>**THERMOMETER** | Cumulative parameterization for an ordinal CLASS variable |
| **POLYNOMIAL**<br>**POLY** | Polynomial coding |
| **REFERENCE**<br>**REF** | Reference cell coding |
| **ORTHEFFECT** | Orthogonalizes PARAM=EFFECT coding |
| **ORTHORDINAL**<br>**ORTHOTHERM** | Orthogonalizes PARAM=ORDINAL coding |
| **ORTHPOLY** | Orthogonalizes PARAM=POLYNOMIAL coding |
| **ORTHREF** | Orthogonalizes PARAM=REFERENCE coding |

All parameterizations are full rank, except for the GLM parameterization. The REF= option in the CLASS statement determines the reference level for EFFECT and REFERENCE coding and for their orthogonal parameterizations. It also indirectly determines the reference level for a singular GLM parameterization through the order of levels.

If a PARAM= option is specified as a variable option for some variables, then any variables for which PARAM= is not specified use either the EFFECT parameterization if the global PARAM= option is not specified, or the full-rank parameterization indicated in the global PARAM= option if specified. If the global PARAM=GLM option is specified and PARAM= is also specified for some variables, GLM parameterization is used for all variables.

If PARAM=ORTHPOLY or PARAM=POLY and the classification variable is numeric, then the ORDER= option in the CLASS statement is ignored, and the internal unformatted values are used. For more information, see the section "Other Parameterizations" on page 416 in Chapter 20, "Shared Concepts and Topics."

**REF=***'level' | keyword*

specifies the reference level for PARAM=EFFECT, PARAM=REFERENCE, and their orthogonalizations. For PARAM=GLM, the REF= option specifies a level of the classification variable to be put at the end of the list of levels. This level thus corresponds to the reference level in the usual interpretation of the linear estimates with a singular parameterization.

For an individual variable REF= option (but not for a global REF= option), you can specify the *level* of the variable to use as the reference level. Specify the formatted value of the variable if a format is assigned. For a global or individual variable REF= option, you can use one of the following *keywords*:

**FIRST** designates the first ordered level as reference.

**LAST** designates the last ordered level as reference.

By default, REF=LAST.

**TRUNCATE***< =n >*

specifies the length *n* of CLASS variable values to use in determining CLASS variable levels. The default is to use the full formatted length of the CLASS variable. If you specify TRUNCATE without the length *n*, the first 16 characters of the formatted values are used. The TRUNCATE option is available only as a global option.

## Class Variable Default Parameterization

If you do not specify the PARAM= option, the default PARAM=GLM parameterization is used.

## Class Variable Naming Convention

Parameter names for a CLASS predictor variable are constructed by concatenating the CLASS variable name with the CLASS levels. However, for the POLYNOMIAL and orthogonal parameterizations, parameter names are formed by concatenating the CLASS variable name and keywords that reflect the parameterization. For examples and more information, see the section "Other Parameterizations" on page 416 in Chapter 20, "Shared Concepts and Topics."

## Class Variable Parameterization with Unbalanced Designs

PROC GENMOD initially parameterizes the CLASS variables by looking at the levels of the variables across the complete data set. If you have an *unbalanced* replication of levels across variables or BY groups, then the design matrix and the parameter interpretation might be different from what you expect. For example, suppose you have a model that has one CLASS variable A with three levels (1, 2, and 3) and another CLASS variable B with two levels (1 and 2). If the third level of A occurs only with the first level of B, if you use the EFFECT parameterization, and if your model contains the effect A(B) and an intercept, then the design for A within the second level of B is not a differential effect. In particular, the design looks like the following:

| | | Design Matrix | | | |
|---|---|---|---|---|---|
| | | A(B=1) | | A(B=2) | |
| **B** | **A** | **A1** | **A2** | **A1** | **A2** |
| 1 | 1 | 1 | 0 | 0 | 0 |
| 1 | 2 | 0 | 1 | 0 | 0 |
| 1 | 3 | –1 | –1 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 | 0 |
| 2 | 2 | 0 | 0 | 0 | 1 |

PROC GENMOD detects linear dependency among the last two design variables and sets the parameter for A2(B=2) to zero, resulting in an interpretation of these parameters as if they were reference- or dummy-coded. The REFERENCE or GLM parameterization might be more appropriate for such problems.

## CODE Statement

**CODE** < *options* > **;**

The CODE statement writes SAS DATA step code for computing predicted values of the fitted model either to a file or to a catalog entry. This code can then be included in a DATA step to score new data.

Table 51.3 summarizes the *options* available in the CODE statement.

**Table 51.3** CODE Statement Options

| Option | Description |
|---|---|
| CATALOG= | Names the catalog entry where the generated code is saved |
| DUMMIES | Retains the dummy variables in the data set |
| ERROR | Computes the error function |
| FILE= | Names the file where the generated code is saved |
| FORMAT= | Specifies the numeric format for the regression coefficients |
| GROUP= | Specifies the group identifier for array names and statement labels |
| IMPUTE | Imputes predicted values for observations with missing or invalid covariates |
| LINESIZE= | Specifies the line size of the generated code |
| LOOKUP= | Specifies the algorithm for looking up CLASS levels |
| RESIDUAL | Computes residuals |

For details about the syntax of the CODE statement, see the section “CODE Statement” on page 424 in Chapter 20, “Shared Concepts and Topics.”

## CONTRAST Statement

**CONTRAST** '*label*' *contrast-specification* < / *options* > **;**

The CONTRAST statement provides a means of obtaining a test of a specified hypothesis concerning the model parameters. This is accomplished by specifying a matrix $\mathbf{L}$ for testing the hypothesis $\mathbf{L}'\boldsymbol{\beta} = 0$. 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 3652 and the section "CLASS Statement" on page 3598. 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 and after the ZEROMODEL statement for zero-inflated models. Statistics for multiple CONTRAST statements are displayed in a single table.

The elements of the CONTRAST statement are as follows:

*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.

*contrast-specification* identifies the effects and their coefficients from which the $\mathbf{L}$ matrix is formed. The *contrast-specification* can be specified in two different ways. The first method applies to all models except the zero-inflated (ZI) distributions (zero-inflated Poisson and zero-inflated negative binomial), and the syntax is:

*effect values* < ,*. . . effect values* >

The second method of specifying a contrast applies only to ZI models, and the syntax is:

*effect values* < ,*. . . effect values* > **@ZERO** *effect values* < ,*. . . effect values* >

where

*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.

*options* specifies CONTRAST statement options.

Specification of sets of *effect values* before the @ZERO separator results in a row of the $\mathbf{L}'$ matrix with coefficients for *effects* in the regression part of the model set to *values* and with the coefficients for the zero-inflation part of the model set to zero. Specification of sets of *effect values* after the @ZERO separator results in a row of the $\mathbf{L}$ matrix with the coefficients for the regression part of the model set to zero and with the coefficients of *effects* in the zero-inflation part of the model set to *values*. For example, the statements

```
class a;
model y=a;
contrast 'Label1' A 1 -1;
```

specify an $\mathbf{L}'$ matrix with one row with coefficients 1 for the first level of A and –1 for the second level of A.

The statements

```
   class a b;
   model y=a / dist=zip;
   zeromodel b;
   contrast 'Label2' A 1 -1 @zero B 1 -1;
```

specify an $\mathbf{L}'$ matrix with two rows: the first row has coefficients 1 for the first level of A, –1 for the second level of A, and zeros for all levels of B; the second row has coefficients 0 for all levels of A, 1 for the first level of B, and –1 for the second level of B.

The rows of $\mathbf{L}'$ are specified in order and are separated by commas.

If you use the default less-than-full-rank PROC GLM CLASS variable parameterization, each row of the $\mathbf{L}'$ matrix is checked for estimability. If PROC GENMOD finds a contrast to be nonestimable, it displays missing values in corresponding rows in the results. See Searle (1971) for a discussion of estimable functions. 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.

When you use any of the full-rank PARAM= CLASS variable options, all parameters are directly estimable, and rows of $\mathbf{L}'$ are not checked for estimability.

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 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 section "Generalized Score Statistics" on page 3671 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*

**EPSILON=***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. Let $\mathbf{K}'$ be any row in the contrast matrix $\mathbf{L}$. Define C to be equal to ABS($\mathbf{K}'$) if ABS($\mathbf{K}'$) is greater than 0; otherwise, C equals 1. 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. The SINGULAR= option in the MODEL statement affects the computation of the generalized inverse of the matrix $\mathbf{X}'\mathbf{X}$. It might also be necessary to adjust this value for some data.

**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 statement 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 3630) 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.

## EFFECTPLOT Statement

**EFFECTPLOT** < *plot-type* < *(plot-definition-options)* > > < / *options* > ;

The EFFECTPLOT statement produces a display of the fitted model and provides options for changing and enhancing the displays. Table 51.4 describes the available *plot-types* and their *plot-definition-options*.

**Table 51.4** *Plot-Types* and *Plot-Definition-Options*

| Plot-Type and Description | Plot-Definition-Options |
|---|---|
| **BOX** | |
| Displays a box plot of continuous response data at each level of a CLASS effect, with predicted values superimposed and connected by a line. This is an alternative to the INTERACTION *plot-type*. | PLOTBY= variable or CLASS effect<br>X= CLASS variable or effect |
| **CONTOUR** | |
| Displays a contour plot of predicted values against two continuous covariates. | PLOTBY= variable or CLASS effect<br>X= continuous variable<br>Y= continuous variable |
| **FIT** | |
| Displays a curve of predicted values versus a continuous variable. | PLOTBY= variable or CLASS effect<br>X= continuous variable |
| **INTERACTION** | |
| Displays a plot of predicted values (possibly with error bars) versus the levels of a CLASS effect. The predicted values are connected with lines and can be grouped by the levels of another CLASS effect. | PLOTBY= variable or CLASS effect<br>SLICEBY= variable or CLASS effect<br>X= CLASS variable or effect |
| **MOSAIC** | |
| Displays a mosaic plot of predicted values using up to three CLASS effects. | PLOTBY= variable or CLASS effect<br>X= CLASS effects |
| **SLICEFIT** | |
| Displays a curve of predicted values versus a continuous variable grouped by the levels of a CLASS effect. | PLOTBY= variable or CLASS effect<br>SLICEBY= variable or CLASS effect<br>X= continuous variable |

For full details about the syntax and options of the EFFECTPLOT statement, see the section “EFFECTPLOT Statement” on page 447 in Chapter 20, “Shared Concepts and Topics.”

## ESTIMATE Statement

**ESTIMATE** '*label*' *contrast-specification* < / *options* > ;

The ESTIMATE statement is similar to a CONTRAST statement, except only one-row $\mathbf{L}'$ matrices are permitted.

The elements of the ESTIMATE statement are as follows:

*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.

*contrast-specification* identifies the effects and their coefficients from which the $\mathbf{L}$ matrix is formed. The *contrast-specification* can be specified in two different ways. The first method applies to all models except the zero-inflated (ZI) distributions (zero-inflated Poisson and zero-inflated negative binomial), and the syntax is:

*effect values* < . . . *effect values* >

The second method of specifying a contrast applies only to ZI models, and the syntax is:

*effect values* < . . . *effect values* > **@ZERO** *effect values* < . . . *effect values* >

where

*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.

*options* specifies options for the ESTIMATE statement.

For ZI models, sets of *effects values* before the @ZERO separator correspond to the regression part of the model with regression parameters $\boldsymbol{\beta}$, and *effects values* after the @ZERO separator correspond to the zero-inflation part of the model with regression parameters $\boldsymbol{\gamma}$. In the case of ZI models, a one-row $\mathbf{L}'$ matrix is created for the regression part of the model, another one-row $\mathbf{L}'$ matrix is created for the zero-inflation part of the model, and separate estimates for the two $\mathbf{L}$ matrices are computed and displayed.

If you use the default less-than-full-rank GLM CLASS variable parameterization, 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. See Searle (1971) for a discussion of estimable functions.

The actual estimate, $\mathbf{L}'\boldsymbol{\beta}$ (and $\mathbf{L}'\boldsymbol{\gamma}$ for ZI models), its approximate standard error, and confidence limits are displayed. Additionally, the corresponding estimate on the mean scale (defined as the inverse link function applied to $\mathbf{L}'\boldsymbol{\beta}$), and confidence limits are displayed. Wald chi-square tests that $\mathbf{L}'\boldsymbol{\beta} = 0$ and $\mathbf{L}'\boldsymbol{\gamma} = 0$ are also displayed.

The approximate standard error of the estimate is computed as the square root of $\mathbf{L}'\hat{\boldsymbol{\Sigma}}\mathbf{L}$, where $\hat{\boldsymbol{\Sigma}}$ 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 and the checking for estimability for an ESTIMATE statement follow 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.

**DIVISOR=***number*

specifies a value by which to divide all coefficients so that fractional coefficients can be entered as integer numerators. For example, you can use

```
estimate '1/3(A1+A2) - 2/3A3' a 1 1 -2 / divisor=3;
```

instead of

```
estimate '1/3(A1+A2) - 2/3A3' a 0.33333 0.33333 -0.66667;
```

**E**

requests that the **L** matrix coefficients be displayed.

**EXP**

requests that $\exp(\mathbf{L}'\boldsymbol{\beta})$, its standard error, and its confidence limits be computed. If you specify the EXP option, standard errors are computed using the delta method. Confidence limits are computed by exponentiating the confidence limits for $\mathbf{L}'\boldsymbol{\beta}$.

**SINGULAR=***number*

**EPSILON=***number*

tunes the estimability checking as described for the CONTRAST statement.

## EXACT Statement

**EXACT** < '*label*' > < **INTERCEPT** > < *effects* > < / *options* > **;**

The EXACT statement performs exact tests of the parameters for the specified *effects* and optionally estimates the parameters and outputs the exact conditional distributions. You can specify the keyword INTERCEPT and any *effects* in the MODEL statement. Inference on the parameters of the specified effects is performed by conditioning on the sufficient statistics of all the other model parameters (possibly including the intercept).

You can specify several EXACT statements, but they must follow the MODEL statement. Each statement can optionally include an identifying *label*. If several EXACT statements are specified, any statement without a label is assigned a label of the form “Exact*n*,” where *n* indicates the *n*th EXACT statement. The label is included in the headers of the displayed exact analysis tables.

If a STRATA statement is also specified, then a stratified exact logistic regression or a stratified exact Poisson regression is performed. The model contains a different intercept for each stratum, and these intercepts are conditioned out of the model along with any other nuisance parameters (parameters for effects specified in the MODEL statement that are not in the EXACT statement).

The ASSESSMENT, BAYES, CONTRAST, EFFECTPLOT, ESTIMATE, LSMEANS, LSMESTIMATE, OUTPUT, SLICE, and STORE statements are not available with an exact analysis. Exact analyses are not performed when you specify a WEIGHT statement, or a model other than LINK=LOGIT with DIST=BIN or LINK=LOG with DIST=POISSON. An OFFSET= variable is not available with exact logistic regression. Exact estimation is not available for ordinal response models.

For classification variables, use of the reference parameterization is recommended.

The following *options* can be specified in each EXACT statement after a slash (/):

**ALPHA=***number*

specifies the level of significance $\alpha$ for $100(1-\alpha)\%$ confidence limits for the parameters or odds ratios. The value of *number* must be between 0 and 1. By default, *number* is equal to the value of the ALPHA= option in the MODEL statement, or 0.05 if that option is not specified.

**CLTYPE=EXACT | MIDP**

requests either the exact or mid-$p$ confidence intervals for the parameter estimates. By default, the exact intervals are produced. The confidence coefficient can be specified with the ALPHA= option. The mid-$p$ interval can be modified with the MIDPFACTOR= option. See the section "Exact Logistic and Exact Poisson Regression" on page 3684 for details.

**ESTIMATE < =***keyword* **>**

estimates the individual parameters (conditioned on all other parameters) for the effects specified in the EXACT statement. For each parameter, a point estimate, a standard error, a confidence interval, and a $p$-value for a two-sided test that the parameter is zero are displayed. Note that the two-sided $p$-value is twice the one-sided $p$-value. You can optionally specify one of the following *keywords*:

**PARM** specifies that the parameters be estimated. This is the default.

**ODDS** specifies that the odds ratios be estimated. If you have classification variables, then you must also specify the PARAM=REF option in the CLASS statement.

**BOTH** specifies that both the parameters and odds ratios be estimated.

**JOINT**

performs the joint test that all of the parameters are simultaneously equal to zero, performs individual hypothesis tests for the parameter of each continuous variable, and performs joint tests for the parameters of each classification variable. The joint test is indicated in the "Conditional Exact Tests" table by the label "Joint."

**JOINTONLY**

performs only the joint test of the parameters. The test is indicated in the "Conditional Exact Tests" table by the label "Joint." When this option is specified, individual tests for the parameters of each continuous variable and joint tests for the parameters of the classification variables are not performed.

**MIDPFACTOR=**$\delta_1$ **|** $(\delta_1, \delta_2)$

sets the tie factors used to produce the mid-$p$ hypothesis statistics and the mid-$p$ confidence intervals. $\delta_1$ modifies both the hypothesis tests and confidence intervals, while $\delta_2$ affects only the hypothesis tests. By default, $\delta_1 = 0.5$ and $\delta_2 = 1.0$. See the section "Exact Logistic and Exact Poisson Regression" on page 3684 for details.

**ONESIDED**

requests one-sided confidence intervals and *p*-values for the individual parameter estimates and odds ratios. The one-sided *p*-value is the smaller of the left- and right-tail probabilities for the observed sufficient statistic of the parameter under the null hypothesis that the parameter is zero. The two-sided *p*-values (default) are twice the one-sided *p*-values. See the section "Exact Logistic and Exact Poisson Regression" on page 3684 for more details.

**OUTDIST=***SAS-data-set*

names the SAS data set that contains the exact conditional distributions. This data set contains all of the exact conditional distributions that are required to process the corresponding EXACT statement. This data set contains the possible sufficient statistics for the parameters of the effects specified in the EXACT statement, the counts, and, when hypothesis tests are performed on the parameters, the probability of occurrence and the score value for each sufficient statistic. When you request an OUTDIST= data set, the observed sufficient statistics are displayed in the "Sufficient Statistics" table. See the section "OUTDIST= Output Data Set" on page 3685 for more information.

## EXACT Statement Examples

In the following example, two exact tests are computed: one for x1 and the other for x2. The test for x1 is based on the exact conditional distribution of the sufficient statistic for the x1 parameter given the observed values of the sufficient statistics for the intercept, x2, and x3 parameters; likewise, the test for x2 is conditional on the observed sufficient statistics for the intercept, x1, and x3.

```
proc genmod;
   model y= x1 x2 x3/d=b;
   exact x1 x2;
run;
```

PROC GENMOD determines, from all the specified EXACT statements, the distinct conditional distributions that need to be evaluated. For example, there is only one exact conditional distribution for the following two EXACT statements:

```
exact 'One' x1 / estimate=parm;
exact 'Two' x1 / estimate=parm onesided;
```

For each EXACT statement, individual tests for the parameters of the specified effects are computed unless the JOINTONLY option is specified. Consider the following EXACT statements:

```
exact 'E12' x1 x2 / estimate;
exact 'E1'  x1    / estimate;
exact 'E2'  x2    / estimate;
exact 'J12' x1 x2 / joint;
```

In the E12 statement, the parameters for x1 and x2 are estimated and tested separately. Specifying the E12 statement is equivalent to specifying both the E1 and E2 statements. In the J12 statement, the joint test for the parameters of x1 and x2 is computed in addition to the individual tests for x1 and x2.

## EXACTOPTIONS Statement

**EXACTOPTIONS** *options* **;**

The EXACTOPTIONS statement specifies *options* that apply to every EXACT statement in the program. The following *options* are available:

**ABSFCONV=***value*

specifies the absolute function convergence criterion. Convergence requires a small change in the log-likelihood function in subsequent iterations,

$$|l_i - l_{i-1}| < \mathit{value}$$

where $l_i$ is the value of the log-likelihood function at iteration $i$.

By default, ABSFCONV=1E–12. You can also specify the FCONV= and XCONV= criteria; optimizations are terminated as soon as one criterion is satisfied.

**ADDTOBS**

adds the observed sufficient statistic to the sampled exact distribution if the statistic was not sampled. This option has no effect unless the METHOD=NETWORKMC option is specified and the ESTIMATE option is specified in the EXACT statement. If the observed statistic has not been sampled, then the parameter estimate does not exist; by specifying this option, you can produce (biased) estimates.

**BUILDSUBSETS**

builds every distribution for sampling. By default, some exact distributions are created by taking a subset of a previously generated exact distribution. When the METHOD=NETWORKMC option is invoked, this subsetting behavior has the effect of using fewer than the desired $n$ samples; see the N= option for more details. Use the BUILDSUBSETS option to suppress this subsetting.

**EPSILON=***value*

controls how the partial sums $\sum_{i=1}^{j} y_i \mathbf{x}_i$ are compared. *value* must be between 0 and 1; by default, *value*=1E–8.

**FCONV=***value*

specifies the relative function convergence criterion. Convergence requires a small relative change in the log-likelihood function in subsequent iterations,

$$\frac{|l_i - l_{i-1}|}{|l_{i-1}| + 1\text{E–}6} < \mathit{value}$$

where $l_i$ is the value of the log likelihood at iteration $i$.

By default, FCONV=1E–8. You can also specify the ABSFCONV= and XCONV= criteria; if you specify more than one criterion, then optimizations are terminated as soon as one criterion is satisfied.

**MAXTIME=***seconds*

specifies the maximum clock time (in seconds) that PROC GENMOD can use to calculate the exact distributions. If the limit is exceeded, the procedure halts all computations and prints a note to the SAS log. The default maximum clock time is seven days.

**METHOD=*keyword***

specifies which exact conditional algorithm to use for every EXACT statement specified. You can specify one of the following *keywords*:

**DIRECT** invokes the multivariate shift algorithm of Hirji, Mehta, and Patel (1987). This method directly builds the exact distribution, but it can require an excessive amount of memory in its intermediate stages. METHOD=DIRECT is invoked by default when you are conditioning out at most the intercept.

**NETWORK** invokes an algorithm described in Mehta, Patel, and Senchaudhuri (1992). This method builds a network for each parameter that you are conditioning out, combines the networks, then uses the multivariate shift algorithm to create the exact distribution. The NETWORK method can be faster and require less memory than the DIRECT method. The NETWORK method is invoked by default for most analyses.

**NETWORKMC** invokes the hybrid network and Monte Carlo algorithm of Mehta, Patel, and Senchaudhuri (1992). This method creates a network, then samples from that network; this method does not reject any of the samples at the cost of using a large amount of memory to create the network. METHOD=NETWORKMC is most useful for producing parameter estimates for problems that are too large for the DIRECT and NETWORK methods to handle and for which asymptotic methods are invalid—for example, for sparse data on a large grid.

**N=*n***

specifies the number of Monte Carlo samples to take when you specify the METHOD=NETWORKMC option. By default, $n$ = 10,000. If PROC GENMOD cannot obtain $n$ samples because of a lack of memory, then a note is printed in the SAS log (the number of valid samples is also reported in the listing) and the analysis continues.

The number of samples used to produce any particular statistic might be smaller than $n$. For example, let *X1* and *X2* be continuous variables, denote their joint distribution by $f(X1,X2)$, and let $f(X1 \mid X2 = x2)$ denote the marginal distribution of *X1* conditioned on the observed value of *X2*. If you request the JOINT test of *X1* and *X2*, then $n$ samples are used to generate the estimate $\hat{f}(X1,X2)$ of $f(X1,X2)$, from which the test is computed. However, the parameter estimate for *X1* is computed from the subset of $\hat{f}(X1,X2)$ that has $X2 = x2$, and this subset need not contain $n$ samples. Similarly, the distribution for each level of a classification variable is created by extracting the appropriate subset from the joint distribution for the CLASS variable.

In some cases, the marginal sample size can be too small to admit accurate estimation of a particular statistic; a note is printed in the SAS log when a marginal sample size is less than 100. Increasing $n$ increases the number of samples used in a marginal distribution; however, if you want to control the sample size exactly, you can either specify the BUILDSUBSETS option or do both of the following:

- Remove the JOINT option from the EXACT statement.
- Create dummy variables in a DATA step to represent the levels of a CLASS variable, and specify them as independent variables in the MODEL statement.

**NOLOGSCALE**

specifies that computations for the exact conditional models be computed by using normal scaling. Log scaling can handle numerically larger problems than normal scaling; however, computations in the log scale are slower than computations in normal scale.

**ONDISK**

uses disk space instead of random access memory to build the exact conditional distribution. Use this option to handle larger problems at the cost of slower processing.

**SEED=*seed***

specifies the initial seed for the random number generator used to take the Monte Carlo samples when you specify the METHOD=NETWORKMC option. The value of the SEED= option must be an integer. If you do not specify a seed, or if you specify a value less than or equal to 0, then PROC GENMOD uses the time of day from the computer's clock to generate an initial seed.

**STATUSN=*number***

prints a status line in the SAS log after every *number* of Monte Carlo samples when you specify the METHOD=NETWORKMC option. The number of samples that are taken and the current exact $p$-value for testing the significance of the model are displayed. You can use this status line to track the progress of the computation of the exact conditional distributions.

**STATUSTIME=*seconds***

specifies the time interval (in seconds) for printing a status line in the SAS log. You can use this status line to track the progress of the computation of the exact conditional distributions. The time interval that you specify is approximate; the actual time interval varies. By default, no status reports are produced.

**XCONV=*value***

specifies the relative parameter convergence criterion. Convergence requires a small relative parameter change in subsequent iterations,

$$\max_j |\delta_j^{(i)}| < value$$

where

$$\delta_j^{(i)} = \begin{cases} \beta_j^{(i)} - \beta_j^{(i-1)} & |\beta_j^{(i-1)}| < 0.01 \\ \frac{\beta_j^{(i)} - \beta_j^{(i-1)}}{\beta_j^{(i-1)}} & \text{otherwise} \end{cases}$$

and $\beta_j^{(i)}$ is the estimate of the $j$th parameter at iteration $i$.

By default, XCONV=1E–4. You can also specify the ABSFCONV= and FCONV= criteria; if more than one criterion is specified, then optimizations are terminated as soon as one criterion is satisfied.

## FREQ Statement

**FREQ** *variable* **;**

**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 missing, the observation is not used. In the case of models fit with generalized estimating equations (GEEs), the frequencies apply to the subject/cluster and therefore must be the same for all observations within each subject.

## 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 section "Programming Statements" on page 3630) 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 by 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 3630) 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** < *model-effects* > < / *options* > **;**

The LSMEANS statement computes and compares least squares means (LS-means) of fixed effects. LS-means are *predicted population margins*—that is, they estimate the marginal means over a balanced population. In a sense, LS-means are to unbalanced designs as class and subclass arithmetic means are to balanced designs.

Table 51.5 summarizes the *options* available in the LSMEANS statement. If you specify the BAYES statement, the ADJUST=, STEPDOWN, and LINES options are ignored. The PLOTS= option is not available for a maximum likelihood analysis; it is available only for a Bayesian analysis.

If you specify a zero-inflated model (that is, a model for either the zero-inflated Poisson or the zero-inflated negative binomial distribution), then the least squares means are computed only for effects in the model for the distribution mean, and not for effects in the zero-inflation probability part of the model.

**Table 51.5** LSMEANS Statement Options

| Option | Description |
|---|---|
| **Construction and Computation of LS-Means** | |
| AT | Modifies the covariate value in computing LS-means |
| BYLEVEL | Computes separate margins |
| DIFF | Computes differences of LS-means |
| OM= | Specifies the weighting scheme for LS-means computation as determined by the input data set |
| SINGULAR= | Tunes estimability checking |
| **Degrees of Freedom and *p*-Values** | |
| ADJUST= | Determines the method of multiple-comparison adjustment of LS-means differences |
| ALPHA=$\alpha$ | Determines the confidence level $(1 - \alpha)$ |
| STEPDOWN | Adjusts multiple-comparison *p*-values further in a step-down fashion |
| **Statistical Output** | |
| ASC | Sorts LS-means in ascending order in results that you display by specifying the LINES option |
| CL | Constructs confidence limits for means and mean differences |
| CORR | Displays the correlation matrix of LS-means |
| COV | Displays the covariance matrix of LS-means |
| E | Prints the **L** matrix |
| LINES | Uses connecting lines to indicate nonsignificantly different subsets of LS-means |
| LINESTABLE | Displays the results of the LINES option as a table |
| MEANS | Prints the LS-means |
| PLOTS= | Produces graphs of means and mean comparisons |
| SEED= | Specifies the seed for computations that depend on random numbers |

**Table 51.5** *continued*

| Option | Description |
|---|---|
| **Generalized Linear Modeling** | |
| EXP | Exponentiates and displays estimates of LS-means or LS-means differences |
| ILINK | Computes and displays estimates and standard errors of LS-means (but not differences) on the inverse linked scale |
| ODDSRATIO | Reports (simple) differences of least squares means in terms of odds ratios if permitted by the link function |

For details about the syntax of the LSMEANS statement, see the section "LSMEANS Statement" on page 492 in Chapter 20, "Shared Concepts and Topics."

## LSMESTIMATE Statement

**LSMESTIMATE** *model-effect* < '*label*' > *values* < *divisor*=*n* >
< **,** . . . < '*label*' > *values* < *divisor*=*n* > >
< / *options* > **;**

The LSMESTIMATE statement provides a mechanism for obtaining custom hypothesis tests among least squares means.

Table 51.6 summarizes the *options* available in the LSMESTIMATE statement.

**Table 51.6** LSMESTIMATE Statement Options

| Option | Description |
|---|---|
| **Construction and Computation of LS-Means** | |
| AT | Modifies covariate values in computing LS-means |
| BYLEVEL | Computes separate margins |
| DIVISOR= | Specifies a list of values to divide the coefficients |
| OM= | Specifies the weighting scheme for LS-means computation as determined by a data set |
| SINGULAR= | Tunes estimability checking |
| **Degrees of Freedom and *p*-Values** | |
| ADJUST= | Determines the method of multiple-comparison adjustment of LS-means differences |
| ALPHA=$\alpha$ | Determines the confidence level $(1 - \alpha)$ |
| LOWER | Performs one-sided, lower-tailed inference |
| STEPDOWN | Adjusts multiple-comparison *p*-values further in a step-down fashion |
| TESTVALUE= | Specifies values under the null hypothesis for tests |
| UPPER | Performs one-sided, upper-tailed inference |

**Table 51.6** *continued*

| Option | Description |
|---|---|
| **Statistical Output** | |
| CL | Constructs confidence limits for means and mean differences |
| CORR | Displays the correlation matrix of LS-means |
| COV | Displays the covariance matrix of LS-means |
| E | Prints the **L** matrix |
| ELSM | Prints the **K** matrix |
| JOINT | Produces a joint *F* or chi-square test for the LS-means and LS-means differences |
| PLOTS= | Produces graphs of means and mean comparisons |
| SEED= | Specifies the seed for computations that depend on random numbers |
| **Generalized Linear Modeling** | |
| CATEGORY= | Specifies how to construct estimable functions for multinomial data |
| EXP | Exponentiates and displays LS-means estimates |
| ILINK | Computes and displays estimates and standard errors of LS-means (but not differences) on the inverse linked scale |

For more information about the syntax of the LSMESTIMATE statement, see the section "LSMESTIMATE Statement" on page 513 in Chapter 20, "Shared Concepts and Topics."

## MODEL Statement

**MODEL** *response* < *(variable-options)* > **=** < *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* can 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 preceding MODEL statements. The response variable can be numeric or character. Variable options specific to the response variable can be specified in parentheses immediately after the response variable. Identifying the event level for binomial responses and ordering of response levels for multinomial responses is critical in these models. You can use the response variable options to do this.

Responses for the Poisson distribution must be all nonnegative, 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 section "Specification of Effects" on page 3651 for more information. Also see Chapter 53, "The GLM Procedure."

## Response Variable Options

**DESCENDING | DESC**

reverses the order of the response categories. If both the DESCENDING and ORDER= options are specified, PROC GENMOD orders the response categories according to the ORDER= option and then reverses that order. For more information, see the section "Response Level Ordering" on page 3687.

**EVENT='*category*' | *keyword***

specifies the event category for the binary response model. PROC GENMOD models the probability of the event category. The EVENT= option has no effect when there are more than two response categories. You can specify the value (formatted if a format is applied) of the event category in quotation marks, or you can specify one of the following *keywords*:

**FIRST**

designates the first-ordered category as the event.

**LAST**

designates the last-ordered category as the event.

By default, EVENT=FIRST.

One of the most common sets of response levels is {0,1}, where 1 represents the event for which the probability is to be modeled. Consider the example where Y takes the values 1 and 0 for event and nonevent, respectively, and Exposure is the explanatory variable. To specify the value 1 as the event category, use the following MODEL statement:

```
model Y(event='1') = Exposure;
```

**ORDER=DATA | FORMATTED | FREQ | INTERNAL**

specifies the sort order for the levels of the response variable. The following table displays the available ORDER= options.

| ORDER= | 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. When ORDER=FORMATTED is in effect for numeric variables for which you have supplied no explicit format, the levels are ordered by their internal values.

For more information about sort order, see the chapter on the SORT procedure in the *Base SAS Procedures Guide* and the discussion of BY-group processing in *SAS Programmers Guide: Essentials*.

**REFERENCE=**'*category*' | *keyword*

**REF=**'*category*' | *keyword*

specifies the reference category for the binary response model. Specifying one response category as the reference is the same as specifying the other response category as the event category. You can specify the value (formatted if a format is applied) of the reference category in quotation marks, or you can specify one of the following *keywords*:

**FIRST** designates the first-ordered category as the reference.

**LAST** designates the last-ordered category as the reference.

By default, REF=LAST.

Table 51.7 summarizes the *options* available in the MODEL statement.

**Table 51.7** MODEL Statement Options

| Option | Description |
|---|---|
| AGGREGATE= | Specifies the subpopulations |
| ALPHA= | Sets the confidence coefficient |
| CICONV= | Sets the convergence criterion for profile likelihood confidence intervals |
| CL | Displays confidence limits for predicted values |
| CODING= | Uses effect coding for all classification variables |
| CONVERGE= | Sets the convergence criterion |
| CONVH= | Sets the relative Hessian convergence criterion |
| CORRB | Displays the parameter estimate correlation matrix |
| COVB | Displays the parameter estimate covariance matrix |
| DESCENDING | Reverses the order of the response categories |
| DIAGNOSTICS | Displays case deletion diagnostic statistics |
| DIST= | Specifies the built-in probability distribution |
| EVENT= | Specifies the event category for the binary response model |
| EXACTMAX | Names a variable used for performing an exact Poisson regression |
| EXPECTED | Computes covariances and associated statistics by using the expected Fisher information matrix |
| ID= | Displays the values of *variable* in the input data set in the OBSTATS table |
| INITIAL= | Sets initial values for parameter estimates |
| INTERCEPT= | Initializes the intercept term |
| ITPRINT | Displays the iteration history for all iterative processes |
| LINK= | Specifies the link function |
| LOGNB | Computes the maximum likelihood estimate and confidence limits of $k$-based $\log(k)$ |
| LRCI | Computes two-sided confidence intervals for the partially likelihood function |

**Table 51.7** *continued*

| Option | Description |
|---|---|
| MAXITER= | Sets the maximum allowable number of iterations for all iterative computation processes |
| NOINT | Requests that no intercept term |
| NOLOGNB | Computes the maximum likelihood estimate and confidence limits of $k$ based on $k$ |
| NOSCALE | Holds the scale parameter fixed |
| OBSTATS | Displays an additional table of statistics |
| OFFSET= | Specifies a variable in the input data set to be used as an offset |
| ORDER= | Specifies the sort order for the levels of the response variable |
| PREDICTED | Displays predicted values and associated statistics |
| REFERENCE= | Specifies the reference category for the binary response model |
| RESIDUALS | Displays residuals and standardized residuals |
| SCALE= | Sets the value used for the scale |
| SCORING= | Computes the Hessian matrix using the Fisher scoring method |
| SINGULAR= | Sets the tolerance for testing singularity |
| TYPE1 | Performs a Type 1 analysis |
| TYPE3 | Computes statistics for Type 3 contrasts |
| WALD | Requests Wald statistics for Type 3 contrasts |
| WALDCI | Computes two-sided Wald confidence intervals |
| XVARS | Includes the regression variables in the OBSTATS table |

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

**AGGREGATE= (***variable-list***) | ***variable*

**AGGREGATE**

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. Specifying the AGGREGATE option is equivalent to specifying the AGGREGATE= option with a variable list that includes all explanatory variables in the MODEL statement. Pearson chi-square and deviance statistics are not computed for multinomial models unless this option is specified.

**ALPHA=***number*

**ALPH=***number*

**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 3655 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).

**CODING=EFFECT | FULLRANK**

specifies that effect coding be used for all classification variables in the model. This is the same as specifying PARAM=EFFECT as a CLASS statement 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 **g** is the gradient vector, **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.

**DIAGNOSTICS**

**INFLUENCE**

requests that case deletion diagnostic statistics be displayed (see the OBSTATS option).

**DIST=***keyword*

**D=***keyword*

**ERROR=***keyword*

**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. Models for data with correlated responses fit by the GEE method are not available for the zero-inflated distributions.

| DIST= | Distribution | Default Link Function |
|---|---|---|
| BINOMIAL \| BIN \| B | Binomial | Logit |
| GAMMA \| GAM \| G | Gamma | Inverse ( power(–1) ) |
| GEOMETRIC \| GEOM | Geometric | Log |
| 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 |
| TWEEDIE< (*Tweedie-options*) > | Tweedie | Log |
| ZIP | Zero-inflated Poisson | Log/logit |
| ZINB | Zero-inflated negative binomial | Log/logit |

You can specify the following *Tweedie-options* when you specify DIST=TWEEDIE.

**INITIALP=***starting-value*
specifies a starting value for iterative estimation of the Tweedie power parameter.

**P=***power-parameter*
specifies a fixed Tweedie power parameter.

**EPSILON=***tolerance*
specifies the tolerance for series approximation of the Tweedie density function.

**OFFSET=***constant-value*
specifies a constant value to be added to the response variable for evaluating the extended quasi-likelihood. By default, OFFSET=0.5.

**NTHREADS=***number*
specifies the number of threads to be used in computation.

**EXACTMAX< =***variable***>**
names a variable to be used for performing an exact Poisson regression. For each observation, the integer part of the EXACTMAX value should be nonnegative and at least as large as the response value. If the EXACTMAX option is specified without a variable, then default values are computed. See the section "Exact Logistic and Exact Poisson Regression" on page 3684 for information about using this option.

**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. This option does not affect the model fitting, only the way in which the covariance matrix is computed (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 classification 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 classification 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 classification variables must take on the same number of levels in each BY group. Otherwise, classification 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* | *number-list*

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. If you specify a multinomial model for ordinal data, you can specify a *number-list* for the multiple intercepts 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. If you perform a Bayesian analysis by specifying the BAYES statement, the iteration history for computing the mode of the posterior distribution is also displayed.

This option might 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.

**LOGNB**

specifies that the maximum likelihood estimate and confidence limits of the negative binomial dispersion parameter $k$ be computed based $\log(k)$. This is the default method used for the negative binomial dispersion parameter, so that specifying no option or specifying the LOGNB option have the same effect. The GENMOD procedure computes the maximum likelihood estimate of $\log(k)$ and computes confidence limits based on the asymptotic normality of $\log(k)$ rather than of $k$. The results are always reported in terms of $k$ rather than of $\log(k)$. This method ensures that the estimate and confidence limits for $k$ are positive. See Meeker and Escobar (1998, p. 163) for details about this method of computing confidence limits.

**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 section “Confidence Intervals for Parameters” on page 3655 for more information about 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.

**NOLOGNB**

specifies that the maximum likelihood estimate and confidence limits of the negative binomial dispersion parameter $k$ be computed based on $k$ rather than $\log(k)$. If this option is not specified, then the GENMOD procedure computes the maximum likelihood estimate of $\log(k)$ and computes confidence limits based on the asymptotic normality of $\log(k)$ rather than of $k$. The results are always reported in terms of $k$ rather than of $\log(k)$. This method ensures that the estimate and confidence limits for $k$ are positive. See Meeker and Escobar (1998, p. 163) for details about this method of computing confidence limits.

**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.

**OBSTATS**

specifies that an additional table of statistics be displayed. Formulas for the statistics are given in the section "Predicted Values of the Mean" on page 3657, the section "Residuals" on page 3658, and the section "Case Deletion Diagnostic Statistics" on page 3675. Residuals and fit diagnostics are not computed for multinomial models.

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 $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$, except when you have a multinomial model, for which it is not included.
- 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 3657 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
- a Cook distance–type statistic for assessing the influence of individual observations on overall model fit
- observation leverage
- DFBETA, defined as an approximation to $\hat{\beta} - \widehat{\beta_{[i]}}$ for each parameter estimate $\hat{\beta}$, where $\widehat{\beta_{[i]}}$ is the parameter estimate with the $i$th observation deleted
- standardized DFBETA, defined as DFBETA, normalized by its standard deviation
- zero inflation probability for zero-inflated models
- the mean of a zero-inflated response

The following additional cluster deletion diagnostic statistics are created and displayed for each cluster if a REPEATED statement is specified:

- a Cook distance–type statistic for assessing the influence of entire clusters on overall model fit
- a studentized Cook distance–type statistic for assessing influence of clusters
- cluster leverage
- cluster DFBETA for assessing the influence of entire clusters on individual parameter estimates
- cluster DFBETA normalized by its standard deviation

If you specify the multinomial distribution, only regression variable values, response values, predicted values, confidence limits for the predicted values, and the linear predictor are displayed in the table. Residuals and other diagnostic statistics are not available for the multinomial distribution.

The RESIDUALS, DIAGNOSTICS | INFLUENCE, 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. Regression variables, response values, predicted values, confidence limits for the predicted values, linear predictor, raw residuals, Pearson residuals for each observation in the input data set are available. Case deletion diagnostic statistics are available for each observation and for each cluster.

**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.

When you perform an exact Poisson regression with an OFFSET= variable but the EXACTMAX= option is not specified, then if $o_i$ is the offset for the $i$th observation, floor(exp($o_i$)) should be greater than or equal to the response value. See the section "Exact Logistic and Exact Poisson Regression" on page 3684 for information about the use of the offset in the exact Poisson model.

**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. Residuals and other diagnostic statistics are not available for the multinomial distribution (see the OBSTATS option).

**SCALE=***number*

**SCALE=PEARSON | P**

**PSCALE**

**SCALE=DEVIANCE | 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 section "Dispersion Parameter" on page 3650 and the section "Overdispersion" on page 3650 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 be computed using the Fisher 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 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. This value also controls the check on estimability for ESTIMATE and CONTRAST statements.

**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 thought to be less 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, residuals, and case deletion diagnostics. Residuals and diagnostic statistics are not computed for multinomial models.

You can also request these statistics with the OBSTATS, PREDICTED, RESIDUALS, DIAGNOSTICS | INFLUENCE, CL, or XVARS option in the MODEL statement. You can then create a SAS data set containing them with ODS OUTPUT commands.

You might prefer to specify the OUTPUT statement for requesting these statistics since the following are true:

- 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 _LEVEL_ variable. Residuals and other diagnostic statistics are not available for the multinomial distribution.

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 that uses 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 for all the statistics, except for the case deletion diagnostics for individual parameter estimates, DFBETA, DFBETAS, DFBETAC, and DFBETACS. You can list variables enclosed in parentheses to correspond to the variables in the model, or you can specify the keyword *_all_*, without parentheses, to include deletion diagnostics for all of the parameters in the model.

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 and deletion diagnostics are not available for the multinomial model with ordinal data. Some of the case deletion diagnostic statistics apply only to models for correlated data specified with a REPEATED statement. If you request these statistics for ordinary generalized linear models, the values of the corresponding variables are set to missing in the output data set. Formulas for the statistics are given in the section "Predicted Values of the Mean" on page 3657, the section "Residuals" on page 3658, and the section "Case Deletion Diagnostic Statistics" on page 3675.

The *keywords* allowed and the statistics they represent are as follows:

**DFBETA | DBETA** represents the effect of deleting an observation on parameter estimates. If you specify the keyword *_all_* after the equal sign, variables named DFBETA_*ParameterName* will be included in the output data set to contain the values of the diagnostic statistic to measure the influence of deleting a single observation on the individual parameter estimates. *ParameterName* is the name of the regression model parameter formed from the input variable names concatenated with the appropriate levels, if classification variables are involved.

**DFBETAS | DBETAS** represents the effect of deleting an observation on standardized parameter estimates. If you specify the keyword *_all_* after the equal sign, variables named

DFBETAS_*ParameterName* will be included in the output data set to contain the values of the diagnostic statistic to measure the influence of deleting a single observation on the individual parameter estimates. *ParameterName* is the name of the regression model parameter formed from the input variable names concatenated with the appropriate levels, if classification variables are involved.

**DOBS | COOKSD** represents the Cook distance–type statistic to measure the influence of deleting a single observation on the overall model fit.

**HESSWGT** represents the diagonal element of the weight matrix used in computing the Hessian matrix.

**H | LEVERAGE** represents the leverage of a single observation.

**LOWER | L** represents the 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** represents the predicted value of the mean of the response or the predicted probability that the response variable is less than or equal to the value of _LEVEL_ if the multinomial model for ordinal data is used (in other words, $\Pr(\mathrm{Y} \leq \_\mathrm{LEVEL}\_)$, where Y is the response variable).

**PZERO** represents the zero-inflation probability for zero-inflated models.

**RESCHI** represents the Pearson (chi) residual for identifying observations that are poorly accounted for by the model.

**RESDEV** represents the deviance residual for identifying poorly fitted observations.

**RESLIK** represents the likelihood residual for identifying poorly fitted observations.

**RESRAW** represents the raw residual for identifying poorly fitted observations.

**STDRESCHI** represents the standardized Pearson (chi) residual for identifying observations that are poorly accounted for by the model.

**STDRESDEV** represents the standardized deviance residual for identifying poorly fitted observations.

**STDXBETA** represents the standard error estimate of XBETA (see the XBETA keyword).

**UPPER | U** represents the upper confidence limit for the predicted value of the mean, or the upper 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** represents the 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 $\mathbf{x}_i'\boldsymbol{\beta}$.

The *keywords* in the following list apply only to models specified with a REPEATED statement, fit by generalized estimating equations (GEEs).

**CLEVERAGE** represents the leverage of a cluster.

**CLUSTER** represents the numerical cluster index, in order of sorted clusters.

**DCLS | CLUSTERCOOKSD** represents the Cook distance–type statistic to measure the influence of deleting an entire cluster on the overall model fit.

**DFBETAC | DBETAC** represents the effect of deleting an entire cluster on parameter estimates. If you specify the keyword *_all_* after the equal sign, variables named DFBETAC_*ParameterName* will be included in the output data set to contain the values of the diagnostic statistic to measure the influence of deleting the cluster on the individual parameter estimates. *ParameterName* is the name of the regression model parameter formed from the input variable names concatenated with the appropriate levels, if classification variables are involved.

**DFBETACS | DBETACS** represents the effect of deleting an entire cluster on normalized parameter estimates. If you specify the keyword *_all_* after the equal sign, variables named DFBETACS_*ParameterName* will be included in the output data set to contain the values of the diagnostic statistic to measure the influence of deleting the cluster on the individual parameter estimates, normalized by their standard errors. *ParameterName* is the name of the regression model parameter formed from the input variable names concatenated with the appropriate levels, if classification variables are involved.

**MCLS | CLUSTERDFIT** represents the studentized Cook distance–type statistic to measure the influence of deleting an entire cluster on the overall model fit.

## 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 by 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 can be referenced in programming statements. The mean, linear predictor, and response are represented by the automatic variables _MEAN_, _XBETA_, and _RESP_, respectively, 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 statements illustrate 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 statements perform the same analysis as the Poisson regression example in the section "Getting Started: GENMOD Procedure" on page 3561.

The statements 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, as follows:

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

These statements are for illustration, and they work 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 by using programming statements and the FWDLINK, INVLINK, DEVIANCE, and VARIANCE statements.

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

The syntax of programming statements used in PROC GENMOD is identical to that used in the NLMIXED procedure and the GLIMMIX procedure (see Chapter 89, "The NLMIXED Procedure," and Chapter 52, "The GLIMMIX Procedure") and the MODEL procedure (see the *SAS/ETS User's Guide*). Most of the programming statements that can be used in the DATA step can also be used in the GENMOD procedure. See *SAS DATA Step Statements: Reference* for a description of SAS programming statements. The following are some commonly used programming statements.

```
ABORT;
ARRAY arrayname < [ dimensions ] > < $ > < variables-and-constants >;
CALL name < (expression < , expression . . . >) >;
DELETE;
DO < variable = expression < TO expression > < BY expression > >
    < , expression < TO expression > < BY expression > > . . .
    < WHILE expression > < UNTIL expression >;
END;
GOTO statement-label;
IF expression;
IF expression THEN program-statement;
              ELSE program-statement;
variable = expression;
variable + expression;
LINK statement-label;
PUT < variable > < = > . . . ;
RETURN;
SELECT < (expression) >;
STOP;
SUBSTR(variable, index, length)= expression;
WHEN (expression)program-statement;
       OTHERWISE program-statement;
```

## 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.

Table 51.8 summarizes the *options* available in the REPEATED statement.

**Table 51.8** REPEATED Statement Options

| Option | Description |
|---|---|
| ALPHAINIT= | Specifies initial values for log odds ratio regression parameters |
| CONVERGE= | Specifies the convergence criterion for GEE parameter estimation |
| CORRB | Displays the estimated correlation matrix |
| CORRW | Displays the estimated working correlation matrix |
| COVB | Displays the estimated covariance matrix |
| ECORRB | Displays the estimated empirical correlation matrix |
| ECOVB | Displays the estimated empirical covariance matrix |
| INITIAL= | Specifies initial values of the regression parameters estimation |
| INTERCEPT= | Specifies either an initial or a fixed value of the intercept |
| LOGOR= | Specifies the regression structure of the log odds ratio |

**Table 51.8** *continued*

| Option | Description |
|---|---|
| MAXITER= | Specifies the maximum number of iterations |
| MCORRB | Displays the estimated model-based correlation matrix |
| MCOVB | Displays the estimated model-based covariance matrix |
| MODELSE | Displays an analysis of parameter estimates table |
| PRINTMLE | Displays an analysis of maximum likelihood parameter estimates table |
| RUPDATE= | Specifies the number of iterations between updates of the working correlation matrix |
| SORTED | Groups by subject and sorts within subject |
| SUBCLUSTER= | Specifies a variable defining subclusters |
| SUBJECT= | Identifies a different subject, or cluster |
| TYPE= | Specifies the working correlation matrix structure |
| V6CORR | Uses the SAS 'Version 6' method of computing normalized Pearson chi-square |
| WITHIN= | Specifies the order of measurements within subjects |
| YPAIR= | Specifies the pairs of responses |
| ZDATA= | Specifies the full **z** matrix |
| ZROW= | Specifies the rows of the **z** matrix |

**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. If you specify an exchangeable working correlation structure with the CORR=EXCH option, the CORRW option is not needed to view the estimated correlation, since a table is printed by default that contains the single estimated correlation.

**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. Table 51.9 displays the log odds ratio structure *keywords* and the corresponding log odds ratio regression structures. See the section "Alternating Logistic Regressions" on page 3667 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 51.9** Log Odds Ratio Regression Structures

| Keyword | Log Odds Ratio Regression Structure |
|---|---|
| **EXCH** | Exchangeable |
| **FULLCLUST** | Fully parameterized clusters |
| **LOGORVAR**(*variable*) | Indicator variable for specifying block effects |
| **NESTK** | $k$-nested |
| **NEST1** | 1-nested |
| **ZFULL** | Fully specified $\mathbf{z}$ matrix specified in ZDATA= data set |
| **ZREP** | Single cluster specification for replicated $\mathbf{z}$ matrix specified in ZDATA= data set |
| **ZREP**(*matrix*) | Single cluster specification for replicated $\mathbf{z}$ 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 that uses model-based standard errors for inference. By default, an "Analysis of Parameter Estimates" table based on empirical standard errors is displayed.

**PRINTMLE**

displays an analysis of maximum likelihood parameter estimates table. The maximum likelihood estimates are not displayed unless this *option* is specified.

**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. This variable must be listed in the CLASS statement.

**TYPE=***correlation-structure keyword*

**CORR=***correlation-structure keyword*

specifies the structure of the working correlation matrix used to model the correlation of the responses from subjects. Table 51.10 displays the correlation structure *keywords* and the corresponding correlation structures. The default working correlation type is the independent (CORR=IND). See the section "Details: GENMOD Procedure" on page 3640 for definitions of the correlation matrix types. You should specify LOGOR= or TYPE= but not both.

**Table 51.10** 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 × 4 correlation matrix with the following 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 SAS ‘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 might 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.

## SLICE Statement

**SLICE** *model-effect* < / *options* > **;**

The SLICE statement provides a general mechanism for performing a partitioned analysis of the LS-means for an interaction. This analysis is also known as an analysis of simple effects.

This statement uses the same *options* as the LSMEANS statement, which are summarized in Table 20.24 in Chapter 20, "Shared Concepts and Topics." For more information about the syntax of the SLICE statement, see the section "SLICE Statement" on page 550 in Chapter 20, "Shared Concepts and Topics."

## STORE Statement

**STORE** < **OUT=** >*item-store-name* < / **LABEL=**'*label*' > **;**

The STORE statement saves the context and results of the statistical analysis. The resulting item store has a binary file format that cannot be modified. The contents of the item store can be processed using the PLM procedure. For more information about the syntax of the STORE statement, see the section "STORE Statement" on page 554 in Chapter 20, "Shared Concepts and Topics."

## STRATA Statement

**STRATA** *variable* < **(***option***)** > . . . < *variable* < **(***option***)** > > < / *options* > **;**

The STRATA statement names the *variables* that define *strata* or *matched sets* to use in *stratified exact* logistic regression of binary response data, or a *stratified exact* Poisson regression of count data. An EXACT statement must also be specified.

Observations that have the same *variable* values are in the same matched set. For a stratified logistic model, you can analyze 1:1, 1:*n*, *m*:*n*, and general $m_i$:$n_i$ matched sets where the number of cases and controls varies across strata. For a stratified Poisson model, you can have any number of observations in each stratum. At least one variable must be specified to invoke the stratified analysis, and the usual unconditional asymptotic analysis is not performed. The stratified logistic model has the form

$$\text{logit}(\pi_{hi}) = \alpha_h + \mathbf{x}_{hi}'\boldsymbol{\beta}$$

where $\pi_{hi}$ is the event probability for the $i$th observation in stratum $h$ with covariates $\mathbf{x}_{hi}$ and where the stratum-specific intercepts $\alpha_h$ are the nuisance parameters that are to be conditioned out.

STRATA variables can also be specified in the MODEL statement as classification or continuous covariates; however, the effects are nondegenerate only when crossed with a nonstratification variable. Specifying several STRATA statements is the same as specifying one STRATA statement that contains all the strata variables. The STRATA variables can be either character or numeric, and the formatted values of the STRATA variables determine the levels. Thus, you can also use formats to group values into levels; see the discussion of the FORMAT procedure in the *Base SAS Procedures Guide*.

The “Strata Summary” table is displayed by default. For an exact logistic regression, it displays the number of strata that have a specific number of events and non-events. For example, if you are analyzing a 1:5 matched study, this table enables you to verify that every stratum in the analysis has exactly one event and five non-events. Strata that contain only events or only non-events are reported in this table, but such strata are uninformative and are not used in the analysis. For an exact Poisson regression, the “Strata Summary” table displays the number of strata that contain a specific number of observations, which enables you to check whether every stratum in the analysis has the same number of observations.

The ASSESSMENT, BAYES, CONTRAST, EFFECTPLOT, ESTIMATE, LSMEANS, LSMESTIMATE, OUTPUT, REPEATED, SLICE, and STORE statements are not available with a STRATA statement. Exact analyses are not performed when you specify a WEIGHT statement, or a model other than LINK=LOGIT with DIST=BIN or LINK=LOG with DIST=POISSON. An OFFSET= variable is not available with exact logistic regression.

The following *option* can be specified for a stratification variable by enclosing the *option* in parentheses after the variable name, or it can be specified globally for all STRATA variables after a slash (/).

**MISSING**

treats missing values (‘.’, ._, .A, . . . , .Z for numeric variables and blanks for character variables) as valid STRATA variable values.

The following strata *options* are also available after the slash:

**CHECKDEPENDENCY=***keyword*

**CHECK=***keyword*

specifies which variables are to be tested for dependency before the analysis is performed. The available *keywords* are as follows:

**NONE** performs no dependence checking. Typically, a message about a singular information matrix is displayed if you have dependent variables. Dependent variables can be identified after the analysis by noting any missing parameter estimates.

**COVARIATES** checks dependence between covariates and an added intercept. Dependent covariates are removed from the analysis. However, covariates that are linear functions of the strata variable might not be removed, which results in a singular information matrix message being displayed in the SAS log. This is the default.

**ALL** checks dependence between all the strata and covariates. This *option* can adversely affect performance if you have a large number of strata.

**NOSUMMARY**

suppresses the display of the “Strata Summary” table.

**INFO**

displays the “Strata Information” table, which includes the stratum number, levels of the STRATA variables that define the stratum, and the total frequency for each stratum. Since the number of strata can be very large, this table is displayed only by request.

## 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 3630. 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.

## ZEROMODEL Statement

**ZEROMODEL** *effects < / options >* **;**

The ZEROMODEL statement enables you to perform zero-inflated Poisson regression or zero-inflated negative binomial regression when those respective distributions are specified by the DIST= option in the MODEL statement. The effects in the ZEROMODEL statement consist of explanatory variables or combinations of variables for the zero-inflation probability regression model in a zero-inflated model. The same effects can be used in both the ZEROMODEL statement and the MODEL statement, or effects can be used in one statement or the other separately. 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 section "Specification of Effects" on page 3651 for more information. Also see Chapter 53, "The GLM Procedure."

You can specify the following *option* in the ZEROMODEL statement after a slash (/).

**LINK=***keyword*

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

| LINK= | Link Function |
|---|---|
| **CLOGLOG** | |
| **CLL** | Complementary log-log |
| **LOGIT** | Logit |
| **PROBIT** | Probit |

If no LINK= option is supplied, the LOGIT link is used. User-defined link functions are not allowed.

# Details: GENMOD Procedure

## Generalized Linear Models Theory

This is a brief introduction to the theory of generalized linear models.

### 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 can 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 $Y$:

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

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

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

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 zero-inflated Poisson and zero-inflated negative binomial distributions are not generalized linear models. However, the zero-inflated distributions are included in PROC GENMOD since they are useful extensions of generalized linear models. See Long (1997) for a discussion of the zero-inflated Poisson and zero-inflated negative binomial distributions. The PROC GENMOD scale parameter and the variance of $Y$ are also shown.

- Normal:

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

- Inverse Gaussian:

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

- Gamma:

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

- Geometric: This is a special case of the negative binomial with $k = 1$.

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

- Negative binomial:

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

- Poisson:

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

- Binomial:

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

- Multinomial:

$$\begin{aligned} f(y_1,y_2,\ldots,y_k) &= \frac{m!}{y_1!y_2!\cdots y_k!}p_1^{y_1}p_2^{y_2}\cdots p_k^{y_k} \\ \phi &= 1 \end{aligned}$$

- Zero-inflated Poisson:

$$\begin{aligned} f(y) &= \left\{\begin{array}{ll} \omega + (1-\omega)e^{-\lambda} & \text{for } y = 0 \\ (1-\omega)\frac{\lambda^y e^{-\lambda}}{y!} & \text{for } y = 1,2,\ldots \end{array}\right. \\ \phi &= 1 \\ \mu = \text{E}(Y) &= (1-\omega)\lambda \\ \text{Var}(Y) &= (1-\omega)\lambda(1+\omega\lambda) \\ &= \mu + \frac{\omega}{1-\omega}\mu^2 \end{aligned}$$

- Zero-inflated negative binomial:

$$
\begin{aligned}
f(y) &= \begin{cases} \omega + (1-\omega)(1+k\lambda)^{-\frac{1}{k}} & \text{for } y = 0 \\ (1-\omega)\frac{\Gamma(y+1/k)}{\Gamma(y+1)\Gamma(1/k)}\frac{(k\lambda)^y}{(1+k\lambda)^{y+1/k}} & \text{for } y = 1, 2, \ldots \end{cases} \\
\phi &= 1 \\
\text{dispersion} &= k \\
\mu = \mathrm{E}(Y) &= (1-\omega)\lambda \\
\mathrm{Var}(Y) &= (1-\omega)\lambda(1+\omega\lambda+k\lambda) \\
&= \mu + \left(\frac{\omega}{1-\omega} + \frac{k}{1-\omega}\right)\mu^2
\end{aligned}
$$

- Tweedie $(1 < p < 2)$:

$$
\begin{aligned}
f(y) &= \begin{cases} e^{-\lambda} & \text{for } y = 0 \\ e^{-y/\gamma}e^{-\lambda}\sum_{n=1}^{\infty}\frac{\gamma^{-n\alpha}}{\Gamma(n\alpha)}y^{n\alpha-1}\frac{\lambda^n}{n!} & \text{for } y > 0 \end{cases} \\
\phi &= \frac{\lambda^{1-p}(\alpha\gamma)^{2-p}}{2-p} \\
\mu = \mathrm{E}(Y) &= \lambda\alpha\gamma \\
\mathrm{Var}(Y) &= \lambda\alpha\gamma^2 + \lambda\alpha^2\gamma^2
\end{aligned}
$$

The negative binomial and the zero-inflated negative binomial distributions contain 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 $Y = events/trials$. The variance function is $V(\mu) = \mu(1-\mu)$, and the binomial trials parameter $n$ is regarded as a weight $w$.

The density function for the Tweedie distribution when $1 < p < 2$ is expressed in terms of the parameters of the compound Poisson distribution. For more information about this representation, see the section "Tweedie Distribution for Generalized Linear Models" on page 3661. For $p > 2$, the Tweedie random variable has positive support and its density function $f(y)$ can be expressed in terms of stable distributions as defined in Hougaard (1986).

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 working 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

For distributions other than the zero-inflated Poisson or zero-inflated negative binomial, 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.

There are two link functions and linear predictors associated with zero-inflated distributions: one for the zero inflation probability $\omega$, and another for the mean parameter $\lambda$. See the section "Zero-Inflated Models" on page 3660 for more details about zero-inflated distributions.

## 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 $\phi$. Zero-inflated log likelihoods are parameterized in terms two parameters, $\lambda$ and $\omega$. The parameter $\omega$ is the zero-inflation probability, and $\lambda$ is a function of the distribution mean. The relationship between the mean of the zero-inflated Poisson and zero-inflated negative binomial distributions and the parameter $\lambda$ is defined in the section "Response Probability Distributions" on page 3640. The term $y_i$ represents the response for the $i$th observation, $f_i$ represents a frequency weight that is specified in a FREQ statement, and $w_i$ represents a known dispersion weight that is specified in a WEIGHT statement. If no WEIGHT statement is specified, then $w_i = 1$ for all observations. If no FREQ statement is specified, then $f_i = 1$ for all observations. The log-likelihood functions are of the form

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

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.

For the discrete distributions (binomial, multinomial, negative binomial, and Poisson), the functions computed as the sum of the $l_i$ terms are not proper log-likelihood functions, since terms involving binomial coefficients or factorials of the observed counts are dropped from the computation of the log likelihood, and a dispersion parameter $\phi$ is included in the computation. Deletion of factorial terms and inclusion of a dispersion parameter do not affect parameter estimates or their estimated covariances for these distributions, and this is the function used in maximum likelihood estimation. The value of $\phi$ used in computing the reported log-likelihood function is either the final estimated value, or the fixed value, if the dispersion parameter is fixed. Even though it is not a proper log-likelihood function in all cases, the function computed as the sum of the $l_i$ terms is reported in the output as the *log likelihood*. The proper log-likelihood function is also computed as the sum of the $ll_i$ terms in the following list, and it is reported as the *full log likelihood* in the output.

- Normal:

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

- Inverse Gaussian:

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

- Gamma:

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

- Negative binomial:

$$l_i = y_i\log\left(\frac{k\mu_i}{w_i}\right) - (y_i + w_i/k)\log\left(1+\frac{k\mu_i}{w_i}\right) + \log\left(\frac{\Gamma(y_i + w_i/k)}{\Gamma(w_i/k)}\right)$$

$$ll_i = y_i\log\left(\frac{k\mu_i}{w_i}\right) - (y_i + w_i/k)\log\left(1+\frac{k\mu_i}{w_i}\right) + \log\left(\frac{\Gamma(y_i + w_i/k)}{\Gamma(y_i+1)\Gamma(w_i/k)}\right)$$

- Poisson:

$$l_i = \frac{w_i}{\phi}[y_i\log(\mu_i) - \mu_i]$$

$$ll_i = w_i[y_i\log(\mu_i) - \mu_i - \log(y_i!)]$$

- Binomial:

$$l_i = \frac{w_i}{\phi}[r_i\log(p_i) + (n_i - r_i)\log(1-p_i)]$$

$$ll_i = w_i[\log\begin{pmatrix} n_i \\ r_i \end{pmatrix} + r_i\log(p_i) + (n_i - r_i)\log(1-p_i)]$$

- Multinomial ($k$ categories):

$$l_i = \frac{w_i}{\phi}\sum_{j=1}^{k} y_{ij}\log(\mu_{ij})$$

$$ll_i = w_i[\log(m_i!) + \sum_{j=1}^{k}(y_{ij}\log(\mu_{ij}) - \log(y_{ij}!))]$$

- Zero-inflated Poisson:

$$l_i = ll_i = \begin{cases} w_i\log[\omega_i + (1-\omega_i)\exp(-\lambda_i)] & y_i = 0 \\ w_i[\log(1-\omega_i) + y_i\log(\lambda_i) - \lambda_i - \log(y_i!)] & y_i > 0 \end{cases}$$

- Zero-inflated negative binomial:

$$l_i = ll_i = \begin{cases} \log[\omega_i + (1-\omega_i)(1 + \frac{k}{w_i}\lambda)^{-\frac{w_i}{k}}] & y_i = 0 \\ \\ \log(1-\omega_i) + y_i \log\left(\frac{k\lambda}{w_i}\right) & \\ -(y_i + \frac{w_i}{k}) \log\left(1 + \frac{k\lambda}{w_i}\right) & \\ + \log\left(\frac{\Gamma(y_i + \frac{w_i}{k})}{\Gamma(y_i+1)\Gamma(\frac{w_i}{k})}\right) & y_i > 0 \end{cases}$$

- Tweedie:

$$l_i = ll_i = \log\left(f(y_i, \mu_i, \phi/\omega_i, p)\right)$$

## 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 section "Response Probability Distributions" on page 3640 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

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

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

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

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 s and $\mathbf{H}$.

If $\eta_i = \mathrm{x}_i'\boldsymbol{\beta}$ is the linear predictor for observation $i$ and $g$ is the link function, then $\eta_i = g(\mu_i)$, so that $\mu_i = g^{-1}(\mathrm{x}_i'\boldsymbol{\beta})$ 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

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

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

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

where

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

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 scoring method of fitting. Either $\mathbf{W}_o$ or $\mathbf{W}_e$ can be used in the update equation. The GENMOD procedure uses Fisher 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

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

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 $\mathbf{\Sigma}$.

The correlation matrix is the normalized covariance matrix. That is, if $\sigma_{ij}$ is an element of $\mathbf{\Sigma}$, 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. The deviance cannot be directly calculated for zero-inflated models. Twice the negative of the log likelihood is reported instead of the proper deviance for the zero-inflated Poisson and zero-inflated negative binomial.

| Distribution | Deviance |
|---|---|
| Normal | $\sum_i f_i w_i (y_i - \mu_i)^2$ |
| Poisson | $2\sum_i f_i w_i \left[ y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i) \right]$ |
| Binomial | $2\sum_i f_i w_i m_i \left[ y_i \log\left(\frac{y_i}{\mu_i}\right) + (1 - y_i)\log\left(\frac{1-y_i}{1-\mu_i}\right) \right]$ |
| Gamma | $2\sum_i f_i w_i \left[ -\log\left(\frac{y_i}{\mu_i}\right) + \frac{y_i - \mu_i}{\mu_i} \right]$ |
| Inverse Gaussian | $\sum_i \frac{f_i w_i (y_i - \mu_i)^2}{\mu_i^2 y_i}$ |
| Multinomial | $\sum_i \sum_j f_i w_i y_{ij} \log\left(\frac{y_{ij}}{p_{ij} m_i}\right)$ |
| Negative binomial | $2\sum_i f_i \left[ y \log(y/\mu) - (y + w_i/k) \log\left(\frac{y + w_i/k}{\mu + w_i/k}\right) \right]$ |
| Zero-inflated Poisson | $-2\sum_i f_i \begin{cases} w_i \log[\omega_i + (1-\omega_i)\exp(-\mu_i)] & y_i = 0 \\ w_i [\log(1-\omega_i) + y_i \log(\mu_i) - \\ \mu_i - \log(y_i!)] & y_i > 0 \end{cases}$ |
| Zero-inflated negative binomial | $-2\sum_i f_i \begin{cases} \log[\omega_i + (1-\omega_i)(1 + \frac{k}{w_i}\lambda)] & y_i = 0 \\ \log(1-\omega_i) + y_i \log\left(\frac{k\lambda}{w_i}\right) - \\ (y_i + \frac{w_i}{k}) \log\left(1 + \frac{k\lambda}{w_i}\right) + \\ \log\left(\frac{\Gamma(y_i + \frac{w_i}{k})}{\Gamma(y_i + 1)\Gamma(\frac{w_i}{k})}\right) & y_i > 0 \end{cases}$ |
| Tweedie | $-2\sum_i f_i w_i \left[ \frac{y_i^{2-p} - (2-p) y_i \mu_i^{1-p} + (1-p)\mu_i^{2-p}}{(1-p)(2-p)} \right]$ |

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

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

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

In the Tweedie case, the unit deviance is based on the quasi-likelihood function. Taking the limit of the deviance as $p \to 1$ and $p \to 2$ reduces to the deviance for the Poisson and gamma distributions, respectively. For more information about the Tweedie distribution, see the section "Tweedie Distribution for Generalized Linear Models" on page 3661.

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 can 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 submodels. 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 $p$ is 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 section "Type 1 Analysis" on page 3653 for more about the estimation of the dispersion parameter.

## Other Fit Statistics

The Akaike information criterion (AIC) is a measure of goodness of model fit that balances model fit against model simplicity. AIC has the form

$$\mathrm{AIC} = -2\mathrm{LL} + 2p$$

where $p$ is the number of parameters estimated in the model, and LL is the log likelihood evaluated at the value of the estimated parameters. An alternative form is the corrected AIC given by

$$\mathrm{AICC} = -2\mathrm{LL} + 2p\frac{n}{n-p-1}$$

where $n$ is the total number of observations used.

The Bayesian information criterion (BIC) is a similar measure. BIC is defined by

$$\mathrm{BIC} = -2\mathrm{LL} + p\log(n)$$

See Akaike (1981, 1979) for details of AIC and BIC. See Simonoff (2003) for a discussion of using AIC, AICC, and BIC with generalized linear models. These criteria are useful in selecting among regression models, with smaller values representing better model fit. PROC GENMOD uses the full log likelihoods defined in the section "Log-Likelihood Functions" on page 3644, with all terms included, for computing all of the criteria.

## 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,<br>starting point at *value* |
| Present (negative binomial) | Not present | *k* fixed at 0 |

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 section "Response Probability Distributions" on page 3640.

## 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 might 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: $V(\mu) = \phi\mu$

An alternative method to allow for overdispersion in the Poisson distribution is to fit a negative binomial distribution, where $V(\mu) = \mu + k\mu^2$, instead of the Poisson. The parameter $k$ can be estimated by maximum likelihood, thus allowing for overdispersion of a specific form. This is different from the multiplicative overdispersion factor $\phi$, which can accommodate many forms of overdispersion.

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 a 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 by using quasi-likelihoods instead of proper log likelihoods. For details on quasi-likelihood functions, see McCullagh and Nelder (1989, Chapter 9), McCullagh (1983); Hardin and Hilbe (2003).

Although the estimate of the dispersion parameter is often used to indicate overdispersion or underdispersion, this estimate might 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. 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 SAS and are specified in a CLASS statement. These might 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 the 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 classification 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 using asterisks to join two or more continuous variables: X1*X2.
- Main effects are specified by writing classification variables by themselves: A, B, C.
- Crossed effects (interactions) are specified by joining two or more classification variables with asterisks: A*B, B*C, A*B*C.
- Nested effects are specified by following a main effect or crossed effect with a classification variable or list of classification variables enclosed in parentheses: B(A), C(B A), A*B(C). In this example, B(A) is “B nested within A.”
- Combinations of continuous and classification variables can be specified in the same way by 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, the expression A | B is equivalent to A B A*B. The effects in the bar operator can be classification 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 no more than two variables: A B C A*B A*C B*C.

## 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. It is important to understand the ordering of classification variable parameters when you use the ESTIMATE or CONTRAST statement. The ordering of these parameters is displayed in the “CLASS Level Information” table and in tables that display the parameter estimates of the fitted model.

When you specify an overparameterized model by using the PARAM=GLM option in the CLASS statement, some columns of $\mathbf{X}$ can be linearly dependent on other columns. For example, when you specify a model that consists of an intercept term and a classification variable, the column that corresponds to any one of the levels of the classification variable is linearly dependent on the other columns of $\mathbf{X}$. 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 that corresponds 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 classification variable are checked for dependencies can be set by the ORDER= option in the PROC GENMOD statement or by the ORDER= option in the CLASS statement. For full-rank parameterizations, the columns of the $\mathbf{X}$ matrix are designed to be linearly independent.

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 that involve classification variables might 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 section "Goodness of Fit" on page 3647, 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 3656 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 estimates are computed under the constraint that the Type III function of the parameters is equal to 0, by 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 might 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 $\mathbf{L}$.

For models that use less-than-full-rank parameterization (as specified by the PARAM=GLM option in the CLASS statement), a Type 3 test of an effect of interest (main effect or interaction) is a test of the Type III estimable functions that are defined for that effect. When the model contains no missing cells, performing the Type 3 test of a main effect corresponds to testing the hypothesis of equal marginal means. For more information about Type III estimable functions, see Chapter 53, "The GLM Procedure," and Chapter 16, "The Four Types of Estimable Functions." Also see Littell, Freund, and Spector (1991).

For models that use full-rank parameterization, all parameters are estimable when there are no missing cells, so it is unnecessary to define estimable functions. The standard test of an effect of interest in this case is the joint test that the values of the parameters associated with that effect are zero. For a model that uses effects parameterization (as specified by the PARAM=EFFECT option in the CLASS statement), performing the joint test for a main effect is equivalent to testing the equality of marginal means. For a model that uses reference parameterization (as specified by the PARAM=REF option in the CLASS statement), performing the joint test is equivalent to testing the equality of cell means at the reference level of the other model effects. For more information about the coding scheme and the associated interpretation of results, see Muller and Fetterman (2002, Chapter 14).

If there is no interaction term, the Type 3 test of an effect for a model that uses GLM parameterization is the same as the joint test of the effect for the model that uses full-rank parameterization. In this situation, the joint test is also called the Type 3 test. For a model that contains an interaction term and no missing cells, the Type 3 test of a component main effect under GLM parameterization is the same as the joint test of the component main effect under effect parameterization. Both test the equality of cell means. But this Type 3 test differs from the joint test under reference parameterization, which tests the equality of cell means at the reference level of the other component main effect. If some cells are missing, you can obtain meaningful tests only by testing a Type III estimable function, so in this case you should use GLM parameterization.

The results of a Type 3 test or a joint test do not depend on the order in which you specify the terms in the MODEL statement.

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 3671 for more information about generalized score statistics. Wald tests are also available with the Wald option in the CONTRAST statement. In this case, the robust covariance matrix estimate is used for $\mathbf{\Sigma}$ in the Wald statistic.

## 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 $100(1-\alpha)$th 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:

$$|l^*(\beta_j) - l_0| \leq \epsilon$$

$$(\mathbf{s} + \lambda\mathbf{r})'\mathbf{H}^{-1}(\mathbf{s} + \lambda\mathbf{r}) \leq \epsilon$$

where $l^*(\beta_j)$, 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 specify 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 submodel 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.

If you set $k = 0$ in a negative binomial model, $s$ is the score statistic of Cameron and Trivedi (1998) for testing for overdispersion in a Poisson model against alternatives of the form $V(\mu) = \mu + k\mu^2$.

See Rao (1973, p. 417) for more 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 $\hat{\boldsymbol{\beta}}$. The robust estimate of the covariance is used for $\boldsymbol{\Sigma}$ in the case of models fit with GEEs.

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 $100p$th 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. Residuals are available for all generalized linear models except multinomial models for ordinal response data, for which residuals are not available. Raw residuals and Pearson residuals are available for models fit with generalized estimating equations (GEEs).

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. You can request raw residuals in an output data set with the keyword RESRAW in the OUTPUT statement.

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

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

You can request Pearson residuals in an output data set with the keyword RESCHI in the OUTPUT statement.

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:

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

You can request deviance residuals in an output data set with the keyword RESDEV in the OUTPUT statement. For more information about the deviance computations, see the section "Goodness of Fit" on page 3647.

The adjusted Pearson, deviance, and likelihood residuals are defined by Agresti (2002); Williams (1987); 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

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

You can request standardized Pearson residuals in an output data set with the keyword STDRESCHI in the OUTPUT statement. The deviance residuals, standardized to have unit asymptotic variance, are given by

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

where $d_i$ is the contribution to the total deviance from observation $i$, and $\mathrm{sign}(y_i - \mu_i)$ is 1 if $y_i - \mu_i$ is positive and –1 if $y_i - \mu_i$ is negative. You can request standardized deviance residuals in an output data set with the keyword STDRESDEV in the OUTPUT statement. The likelihood residuals are defined by

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

You can request likelihood residuals in an output data set with the keyword RESLIK in the OUTPUT statement.

## Multinomial Models

This type of model applies to cases where an observation can fall into one of $k$ categories. Binary data occur 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 $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$ can be modeled by the multinomial distribution, defined in the section "Response Probability Distributions" on page 3640, with $\sum_j y_{ij} = m_i$. The multinomial model is an *ordinal* model if the categories have a natural order.

Residuals are not available in the OBSTATS table or the output data set for multinomial models.

By default, and consistently with binomial models, the GENMOD procedure orders the response categories for ordinal multinomial models from lowest to highest and models the probabilities of the lower response levels. You can change the way PROC GENMOD orders the response levels with the RORDER= option in the PROC GENMOD statement. The order that PROC GENMOD uses is shown in the "Response Profiles" output table described in the section "Response Profile" on page 3689.

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

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

where $\mu_1, \mu_2, \ldots, \mu_{k-1}$ 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,

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

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

PROC GENMOD estimates the intercept parameters $\mu_1, \mu_2, \ldots, \mu_{k-1}$ 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 option.

## Zero-Inflated Models

Count data that have an incidence of zeros greater than expected for the underlying probability distribution of counts can be modeled with a zero-inflated distribution. In GENMOD, the underlying distribution can be either Poisson or negative binomial. See Lambert (1992), Long (1997) and Cameron and Trivedi (1998) for more information about zero-inflated models. The population is considered to consist of two types of individuals. The first type gives Poisson or negative binomial distributed counts, which might contain zeros. The second type always gives a zero count. Let $\lambda$ be the underlying distribution mean and $\omega$ be the probability of an individual being of the second type. The parameter $\omega$ is called here the *zero-inflation probability*, and is the probability of zero counts in excess of the frequency predicted by the underlying distribution. You can request that the zero inflation probability be displayed in an output data set with the PZERO keyword. The probability distribution of a zero-inflated Poisson random variable $Y$ is given by

$$\Pr(Y = y) = \begin{cases} \omega + (1-\omega)\mathrm{e}^{-\lambda} & \text{for } y = 0 \\ (1-\omega)\frac{\lambda^{y}\mathrm{e}^{-\lambda}}{y!} & \text{for } y = 1, 2, \ldots \end{cases}$$

and the probability distribution of a zero-inflated negative binomial random variable $Y$ is given by

$$\Pr(Y = y) = \begin{cases} \omega + (1-\omega)(1+k\lambda)^{-\frac{1}{k}} & \text{for } y = 0 \\ (1-\omega)\frac{\Gamma(y+1/k)}{\Gamma(y+1)\Gamma(1/k)}\frac{(k\lambda)^{y}}{(1+k\lambda)^{y+1/k}} & \text{for } y = 1, 2, \ldots \end{cases}$$

where $k$ is the negative binomial dispersion parameter.

You can model the parameters $\omega$ and $\lambda$ in GENMOD with the regression models:

$$\begin{aligned} h(\omega_i) &= \mathbf{z}_i'\boldsymbol{\gamma} \\ g(\lambda_i) &= \mathbf{x}_i'\boldsymbol{\beta} \end{aligned}$$

where $h$ is one of the binary link functions: logit, probit, or complementary log-log. The link function $h$ is the logit link by default, or the link function option specified in the ZEROMODEL statement. The link function $g$ is the log link function by default, or the link function specified in the MODEL statement, for both the Poisson and the negative binomial. The covariates $\mathbf{z}_i$ for observation $i$ are determined by the model specified in the ZEROMODEL statement, and the covariates $\mathbf{x}_i$ are determined by the model specified in the MODEL statement. The regression parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are estimated by maximum likelihood.

The mean and variance of $Y$ for the zero-inflated Poisson are given by

$$\begin{aligned} \mathrm{E}(Y) &= \mu = (1-\omega)\lambda \\ \mathrm{Var}(Y) &= \mu + \frac{\omega}{1-\omega}\mu^2 \end{aligned}$$

and for the zero-inflated negative binomial by

$$\begin{aligned} \mathrm{E}(Y) &= \mu = (1-\omega)\lambda \\ \mathrm{Var}(Y) &= \mu + \left(\frac{\omega}{1-\omega} + \frac{k}{1-\omega}\right)\mu^2 \end{aligned}$$

You can request that the mean of $Y$ be displayed for each observation in an output data set with the PRED keyword.

## Tweedie Distribution for Generalized Linear Models

The Tweedie (1984) distribution has nonnegative support and can have a discrete mass at zero, making it useful to model responses that are a mixture of zeros and positive values. The Tweedie distribution belongs to the exponential family, so it conveniently fits in the generalized linear models framework. According to such parameterization, the mean and variance for the Tweedie random variable are $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \phi\mu^p$, respectively, where $\phi$ is the dispersion parameter and $p$ is an extra parameter that controls the variance of the distribution.

The Tweedie family of distributions includes several important distributions for generalized linear models. When $p = 0$, the Tweedie distribution degenerates to the normal distribution; when $p = 1$, it becomes a Poisson distribution; when $p = 2$, it becomes a gamma distribution; when $p = 3$, it is an inverse Gaussian distribution.

Except for these special cases, the probability density function for the Tweedie distribution does not have a closed form and can at best be expressed in terms of series. Numerical approximations are needed to evaluate the density function. Dunn and Smyth (2005) propose using a finite series and provide a formula to determine its lower and upper indices in order to achieve a desired accuracy. Alternatively, you can apply the Fourier transformation on the characteristic function (Dunn and Smyth 2008). These approximations tend to be expensive when a high level of accuracy is demanded or the data volume becomes large. PROC GENMOD uses the series method unless it becomes complicated to do so. In this case, the method that is based on the Fourier transformation is used. The accuracy of approximation is controlled by the EPSILON= option, whose default value is $10^{-5}$.

The Tweedie distribution is not defined when $p$ is between 0 and 1. In practice, the most interesting range is from 1 to 2 in which the Tweedie distribution gradually loses its mass at 0 as it shifts from a Poisson distribution to a gamma distribution. In this case, the Tweedie random variable $Y$ can be generated from a compound Poisson distribution (Smyth 1996) as

$$\begin{aligned} Y &= \Sigma_{i=1}^{T} X_i \\ T &\sim \mathrm{Poisson}(\lambda) \\ X_i &\sim \mathrm{gamma}(\alpha, \gamma) \end{aligned}$$

where $Y = 0$ if $T = 0$, $T$ and $X_i$ are statistically independent, and $\mathrm{gamma}(\alpha, \gamma)$ denotes a gamma random variable that has mean $\alpha\gamma$ and variance $\alpha\gamma^2$. These parameters are determined by the Tweedie parameters as follows:

$$\begin{aligned} \lambda &= \frac{\mu^{2-p}}{\phi(2-p)} \\ \alpha &= \frac{2-p}{p-1} \\ \gamma &= \phi(p-1)\mu^{p-1} \end{aligned}$$

Inversely, given the Tweedie distributional parameters, the parameters of the compound Poisson distribution are determined as follows:

$$\begin{aligned}\mu &= \lambda\alpha\gamma \\ p &= \frac{\alpha+2}{\alpha+1} \\ \phi &= \frac{\lambda^{1-p}(\alpha\gamma)^{2-p}}{2-p}\end{aligned}$$

In terms of generalized linear models parameterizations, the canonical parameter $\theta$ for the Tweedie density can be expressed as

$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p} & p \neq 1 \\ \log\mu & p = 1 \end{cases}$$

and the function $b(\theta)$ is

$$b(\theta) = \begin{cases} \frac{\mu^{2-p}}{2-p} & p \neq 2 \\ \log\mu & p = 2 \end{cases}$$

Because of the intractability of differentiating the gradient functions with respect to the variance parameters, PROC GENMOD uses a quasi-Newton approach to maximize the likelihood function, where the Hessian matrix is approximated by taking finite differences of the gradient functions. Convergence is determined by a union of two criteria: the relative gradient convergence criterion is set to $10^{-9}$, and the relative function convergence criterion is set to $2 \times 10^{-9}$. Convergence is declared when at least one of the criteria is attained during the quasi-Newton iteration.

Before PROC GENMOD maximizes the approximate likelihood, it first maximizes the following extended log quasi-likelihood which is constructed according to the definition of McCullagh and Nelder (1989, Chapter 9) as

$$Q_p(\mathbf{y}, \boldsymbol{\mu}, \phi, p) = \sum_i q(y_i, \mu_i, \phi, p)$$

where the contribution from an observation is

$$q(y_i, \mu_i, \phi, p) = -0.5\log(2\pi\phi y_i^p / w_i) - w_i \left( \frac{y_i^{2-p} - (2-p)y_i\mu_i^{1-p} + (1-p)\mu_i^{2-p}}{(1-p)(2-p)} \right) / \phi$$

and $w_i$ is the weight for the observation from the WEIGHT statement.

The range of parameter $p$ for the quasi-likelihood is from 1 to 2. For a specified P= value outside this range, PROC GENMOD skips optimization of the quasi-likelihood. To maintain numerical stability, PROC GENMOD imposes a lower bound of 1.1 and a upper bound of 1.99 for computation with the quasi-likelihood. The full-likelihood solution imposes the same lower bound but no upper bound. The estimates that are obtained from optimizing the quasi-likelihood are usually near the full-likelihood solution so that fewer iterations are needed for maximizing the more expensive full likelihood.

## 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 $\sum_{i=1}^{K} f_i n_i$ total measurements where $f_i$ is the cluster frequency that you specify in the FREQUENCY statement. If you do not specify a FREQUENCY statement, $f_i = 1$ for all observations. The frequencies must be the same for all observations within each cluster.

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. In the rest of this section, $v(\mu)$ is the variance function of the specified distribution. For the case of the negative binomial, the variance function is fixed at $v(\mu) = \mu + k\mu^2$, where $k$ is either the maximum likelihood estimate of the negative binomial dispersion parameter or the value specified in the NOSCALE and SCALE= options in the MODEL statement.

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

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

where

$$\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$$

Because

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

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

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

## 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

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

where $\mathbf{A}_i$ is an $n_i \times n_i$ diagonal matrix with $v(\mu_{ij})$ as the $j$th diagonal element and $\mathbf{W}_i$ is an $n_i \times n_i$ diagonal matrix with $w_{ij}$ as the $j$th diagonal, where $w_{ij}$ is a weight specified with the WEIGHT statement. If there is no WEIGHT statement, $w_{ij} = 1$ for all $i$ and $j$. 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 by using the current value of the parameter vector $\boldsymbol{\beta}$ to compute appropriate functions of the Pearson residual

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

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** | |
| $\text{Corr}(Y_{ij}, Y_{ik}) = r_{jk}$<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** | |
| $\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}$ | The working correlation is not estimated in this case. |
| ***m*-dependent** | |
| $\text{Corr}(Y_{ij}, Y_{i,j+t}) = \begin{cases} 1 & t = 0 \\ \alpha_t & t = 1, 2, \ldots, m \\ 0 & t > m \end{cases}$ | $\hat{\alpha}_t = \frac{1}{(K_t - p)\phi} \sum_{i=1}^{K} f_i \sum_{j \le n_i - t} e_{ij} e_{i,j+t}$<br>$K_t = \sum_{i=1}^{K} f_i (n_i - t)$ |
| **Exchangeable** | |
| $\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha & j \neq k \end{cases}$ | $\hat{\alpha} = \frac{1}{(N^* - p)\phi} \sum_{i=1}^{K} f_i \sum_{j<k} e_{ij} e_{ik}$<br>$N^* = 0.5 \sum_{i=1}^{K} f_i n_i (n_i - 1)$ |
| **Unstructured** | |
| $\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha_{jk} & j \neq k \end{cases}$ | $\hat{\alpha}_{jk} = \frac{1}{(K - p)\phi} \sum_{i=1}^{K} f_i e_{ij} e_{ik}$ |

**Table 51.11** *continued*

| Working Correlation Structure | Estimator |
|---|---|
| **Autoregressive**<br>**AR(1)**<br>$\text{Corr}(Y_{ij}, Y_{i,j+t}) = \alpha^t$<br>for $t = 0, 1, 2, \ldots, n_i - j$ | $\hat{\alpha} = \frac{1}{(K_1 - p)\phi} \sum_{i=1}^{K} f_i \sum_{j \le n_i - 1} e_{ij} e_{i,j+1}$<br>$K_1 = \sum_{i=1}^{K} f_i(n_i - 1)$ |

## Dispersion Parameter

The dispersion parameter $\phi$ is estimated by

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

where $N = \sum_{i=1}^{K} f_i n_i$ 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. If a fixed scale parameter is specified with the NOSCALE option in the MODEL statement, then the fixed value is used in estimating the model-based covariance matrix and standard errors.

## Fitting Algorithm

The following is an algorithm for fitting the specified model by 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:

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

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

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

5. Repeat steps 2-4 until convergence.

## Missing Data

See 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 by 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. The resulting covariances and standard errors are valid under the missing completely at random (MCAR) assumption.

For example, for the unstructured working correlation model,

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

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 $(\mathbf{Y}_i - \boldsymbol{\mu}_\mathbf{i})$, 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

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

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

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

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 $\mathrm{Cov}(\mathbf{Y}_i) \neq \mathbf{V}_i$. For further information about the robust variance estimate, see Zeger, Liang, and Albert (1988); Royall (1986); White (1982). In computing $\boldsymbol{\Sigma}_e$, $\boldsymbol{\beta}$ 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,

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

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

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

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

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

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, $\gamma_{ijk} = \log(\mathrm{OR}(Y_{ij}, Y_{ik}))$, as

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

where $\boldsymbol{\alpha}$ 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 that uses 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, $\boldsymbol{\alpha}$, 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 $\mathbf{z}$ 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** specifies 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** specifies 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 $\boldsymbol{\alpha}$. For example, if a full cluster is of size 4, then there are $\frac{4\times 3}{2} = 6$ parameters, and the $\mathbf{z}$ matrix is of the form

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

The elements of $\boldsymbol{\alpha}$ 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*) specifies 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** specifies $k$-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**

specifies 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 $\mathbf{z}$ matrix. You must also specify a SAS data set containing the $\mathbf{z}$ matrix with the ZDATA=*data-set-name* option. Each observation in the data set corresponds to one row of the $\mathbf{z}$ 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 $1, 2, \ldots, n$, then the rows of the $\mathbf{z}$ matrix must be specified for pairs in the order $(1,2), (1,3), \ldots, (1,n), (2,3), \ldots, (2,n), \ldots, (n-1,n)$. 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 $\mathbf{z}$ 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 $\mathbf{z}$ matrix with the YPAIR=(*variable1, variable2*) option. If you specify this option, the data from the ZDATA data set are sorted within each cluster by *variable1* and *variable2*. See Example 51.6 for an example of specifying a full $\mathbf{z}$ matrix.

**ZREP**

specifies a replicated $\mathbf{z}$ matrix. You specify $\mathbf{z}$ matrix data exactly as you do for the ZFULL case, except that you specify only one complete cluster. The $\mathbf{z}$ 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).

**ZREP**(*matrix*)

specifies direct input of the replicated $\mathbf{z}$ matrix. You specify the $\mathbf{z}$ 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 $z_1, z_2, \ldots, z_q$ make up the corresponding row of the $\mathbf{z}$ 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 $\mathbf{z}$ matrix for a cluster of size 4 with $q = 2$ log odds ratio parameters. The log odds ratio for the pairs (1 2), (1 3), (1 4) is $\alpha_1$, and the log odds ratio for the pairs (2 3), (2 4), (3 4) is $\alpha_1 + \alpha_2$.

## Quasi-likelihood Information Criterion

The quasi-likelihood information criterion (QIC) was developed by Pan (2001) as a modification of the Akaike information criterion (AIC) to apply to models fit by GEEs.

Define the quasi-likelihood under the independence working correlation assumption, evaluated with the parameter estimates under the working correlation of interest as

$$Q(\hat{\boldsymbol{\beta}}(R), \phi) = \sum_{i=1}^{K} f_i \sum_{j=1}^{n_i} Q(\hat{\boldsymbol{\beta}}(R), \phi; (Y_{ij}, \mathbf{X}_{ij}))$$

where the quasi-likelihood contribution of the $j$th observation in the $i$th cluster is defined in the section "Quasi-likelihood Functions" on page 3670 and $\hat{\boldsymbol{\beta}}(R)$ are the parameter estimates obtained from GEEs with the working correlation of interest $R$.

QIC is defined as

$$\mathrm{QIC}(R) = -2Q(\hat{\boldsymbol{\beta}}(R), \phi) + 2\mathrm{trace}(\hat{\Omega}_I \hat{V}_R)$$

where $\hat{V}_R$ is the robust covariance estimate and $\hat{\Omega}_I$ is the inverse of the model-based covariance estimate under the independent working correlation assumption, evaluated at $\hat{\boldsymbol{\beta}}(R)$, the parameter estimates obtained from GEEs with the working correlation of interest $R$.

PROC GENMOD also computes an approximation to $\mathrm{QIC}(R)$ defined by Pan (2001) as

$$\mathrm{QIC}_u(R) = -2Q(\hat{\boldsymbol{\beta}}(R), \phi) + 2p$$

where $p$ is the number of regression parameters.

Pan (2001) notes that QIC is appropriate for selecting regression models and working correlations, whereas $\mathrm{QIC}_u$ is appropriate only for selecting regression models.

## Quasi-likelihood Functions

See McCullagh and Nelder (1989) and Hardin and Hilbe (2003) for discussions of quasi-likelihood functions. The contribution of observation $j$ in cluster $i$ to the quasi-likelihood function evaluated at the regression parameters $\boldsymbol{\beta}$ is given by $Q(\boldsymbol{\beta}, \phi; (Y_{ij}, \mathbf{X}_{ij})) = \frac{Q_{ij}}{\phi}$, where $Q_{ij}$ is defined in the following list. These are used in the computation of the quasi-likelihood information criteria (QIC) for goodness of fit of models fit with GEEs. The $w_{ij}$ are prior weights, if any, specified with the WEIGHT or FREQ statements. Note that the definition of the quasi-likelihood for the negative binomial differs from that given in McCullagh and Nelder (1989). The definition used here allows the negative binomial quasi-likelihood to approach the Poisson as $k \to 0$.

- Normal:

$$Q_{ij} = -\frac{1}{2} w_{ij} (y_{ij} - \mu_{ij})^2$$

- Inverse Gaussian:

$$Q_{ij} = \frac{w_{ij}(\mu_{ij} - .5y_{ij})}{\mu_{ij}^2}$$

- Gamma:

$$Q_{ij} = -w_{ij}\left[\frac{y_{ij}}{\mu_{ij}} + \log(\mu_{ij})\right]$$

- Negative binomial:

$$Q_{ij} = w_{ij}\left[\log\Gamma\left(y_{ij} + \frac{1}{k}\right) - \log\Gamma\left(\frac{1}{k}\right) + y_{ij}\log\left(\frac{k\mu_{ij}}{1+k\mu_{ij}}\right) + \frac{1}{k}\log\left(\frac{1}{1+k\mu_{ij}}\right)\right]$$

- Poisson:

$$Q_{ij} = w_{ij}(y_{ij}\log(\mu_{ij}) - \mu_{ij})$$

- Binomial:

$$Q_{ij} = w_{ij}[r_{ij}\log(p_{ij}) + (n_{ij} - r_{ij})\log(1 - p_{ij})]$$

- Multinomial ($s$ categories):

$$Q_{ij} = w_{ij}\sum_{k=1}^{s} y_{ijk}\log(\mu_{ijk})$$

## 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 $\mathrm{S}(\tilde{\boldsymbol{\beta}})$ be the generalized estimating equation values at $\tilde{\boldsymbol{\beta}}$.

The generalized score statistic is

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

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.

The preceding development for score tests assumes that the rank of the empirical covariance matrix $\boldsymbol{\Sigma}_e$ is not less than the row rank of the contrast matrix $\mathbf{L}$. When the rank of $\boldsymbol{\Sigma}_e$ is less than the row rank of $\mathbf{L}$, estimability of the function is not sufficient to ensure that the chi-square test statistic has a unique value no matter what kind of generalized inverse is used to compute $(\mathbf{L}'\boldsymbol{\Sigma}_e\mathbf{L})^{-}$.

Although it is extremely rare, it is possible in practice that the uniqueness condition is not satisfied. For example, if the number of clusters is less than the number of nonsingular parameters in the model, then the matrix of coefficients for testing the overall null does not satisfy the uniqueness condition. If this condition is not satisfied, then the chi-square statistic for testing $H\colon \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ is not invariant to the choice of the $g_2$-inverse of $\mathbf{L}\boldsymbol{\Sigma}_e\mathbf{L}'$. This chi-square test is not recommended when the uniqueness condition is not satisfied. An alternative approach would be to increase the number of clusters or to find a parsimonious model so that the number of parameters is less than the number of clusters. When the rank of $\boldsymbol{\Sigma}_e$ is less than the row rank of $\mathbf{L}$ for a test, the procedure prints a note to the SAS log.

## Assessment of Models Based on Aggregates of Residuals

Lin, Wei, and Ying (2002) present graphical and numerical methods for model assessment based on the cumulative sums of residuals over certain coordinates (such as covariates or linear predictors) or some related aggregates of residuals. The distributions of these stochastic processes under the assumed model can be approximated by the distributions of certain zero-mean Gaussian processes whose realizations can be generated by simulation. Each observed residual pattern can then be compared, both graphically and numerically, with a number of realizations from the null distribution. Such comparisons enable you to assess objectively whether the observed residual pattern reflects anything beyond random fluctuation. These procedures are useful in determining appropriate functional forms of covariates and link function. You use the ASSESS|ASSESSMENT statement to perform this kind of model-checking with cumulative sums of residuals, moving sums of residuals, or LOESS smoothed residuals. See Example 51.8 and Example 51.9 for examples of model assessment.

Let the model for the mean be

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

where $\mu_i$ is the mean of the response $y_i$ and $\mathbf{x}_i$ is the vector of covariates for the *i*th observation. Denote the raw residual resulting from fitting the model as

$$e_i = y_i - \hat{\mu}_i$$

and let $x_{ij}$ be the value of the *j*th covariate in the model for observation *i*. Then to check the functional form of the *j*th covariate, consider the cumulative sum of residuals with respect to $x_{ij}$,

$$W_j(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(x_{ij} \le x) e_i$$

where $I()$ is the indicator function. For any *x*, $W_j(x)$ is the sum of the residuals with values of $x_j$ less than or equal to *x*.

Denote the score, or gradient vector, by

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} h(\mathbf{x}'\boldsymbol{\beta})\mathbf{x}_i(y_i - \nu(\mathbf{x}'\boldsymbol{\beta}))$$

where $\nu(r) = g^{-1}(r)$, and

$$h(r) = \frac{1}{g\prime(\nu(r))V(\nu(r))}$$

Let $\mathbf{J}$ be the Fisher information matrix

$$\mathbf{J}(\boldsymbol{\beta}) = -\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$$

Define

$$\hat{W}_j(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [I(x_{ij} \le x) + \boldsymbol{\eta}'(x; \hat{\boldsymbol{\beta}})\mathbf{J}^{-1}(\hat{\boldsymbol{\beta}})\mathbf{x}_i h(\mathbf{x}'\hat{\boldsymbol{\beta}})] e_i Z_i$$

where

$$\boldsymbol{\eta}(x;\boldsymbol{\beta}) = -\sum_{i=1}^{n} I(x_{ij} \le x)\frac{\partial \nu(\mathbf{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

and $Z_i$ are independent $N(0,1)$ random variables. Then the conditional distribution of $\hat{W}_j(x)$, given $(y_i, \mathbf{x}_i), i = 1, \ldots, n$, under the null hypothesis $H_0$ that the model for the mean is correct, is the same asymptotically as $n \to \infty$ as the unconditional distribution of $W_j(x)$ (Lin, Wei, and Ying 2002).

You can approximate realizations from the null hypothesis distribution of $W_j(x)$ by repeatedly generating normal samples $Z_i, i = 1, \ldots, n$, while holding $(y_i, \mathbf{x}_i), i = 1, \ldots, n$, at their observed values and computing $\hat{W}_j(x)$ for each sample.

You can assess the functional form of covariate *j* by plotting a few realizations of $\hat{W}_j(x)$ on the same plot as the observed $W_j(x)$ and visually comparing to see how typical the observed $W_j(x)$ is of the null distribution samples.

You can supplement the graphical inspection method with a Kolmogorov-type supremum test. Let $s_j$ be the observed value of $S_j = \sup_x |W_j(x)|$. The *p*-value $\Pr[S_j \ge s_j]$ is approximated by $\Pr[\hat{S}_j \ge s_j]$, where $\hat{S}_j = \sup_x |\hat{W}_j(x)|$. $\Pr[\hat{S}_j \ge s_j]$ is estimated by generating realizations of $\hat{W}_j(.)$ (1,000 is the default number of realizations).

You can check the link function instead of the *j*th covariate by using values of the linear predictor $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ in place of values of the *j*th covariate $x_{ij}$. The graphical and numerical methods described previously are then sensitive to inadequacies in the link function.

An alternative aggregate of residuals is the moving sum statistic

$$W_j(x,b) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(x - b \le x_{ij} \le x)e_i$$

If you specify the keyword WINDOW(*b*), then the moving sum statistic with window size *b* is used instead of the cumulative sum of residuals, with $I(x - b \le x_{ij} \le x)$ replacing $I(x_{ij} \le x)$ in the earlier equation.

If you specify the keyword LOESS(*f*), loess smoothed residuals are used in the preceding formulas, where *f* is the fraction of the data to be used at a given point. If *f* is not specified, $f = \frac{1}{3}$ is used. For data $(Y_i, X_i), i = 1, \ldots, n$, define *r* as the nearest integer to $nf$ and *h* as the *r*th smallest among $|X_i - x|, i = 1, \ldots, n$. Let

$$K_i(x) = K\left(\frac{X_i - x}{h}\right)$$

where

$$K(t) = \frac{70}{81}(1 - |t|^3)^3 I(-1 \le t \le 1)$$

Define

$$w_i(x) = K_i(x)[S_2(x) - (X_i - x)S_1(x)]$$

where

$$S_1(x) = \sum_{i=1}^{n} K_i(x)(X_i - x)$$

$$S_2(x) = \sum_{i=1}^{n} K_i(x)(X_i - x)^2$$

Then the loess estimate of $Y$ at $x$ is defined by

$$\hat{Y}(x) = \sum_{i=1}^{n} \frac{w_i(x)}{\sum_{i=1}^{n} w_i(x)} Y_i$$

Loess smoothed residuals for checking the functional form of the $j$th covariate are defined by replacing $Y_i$ with $e_i$ and $X_i$ with $x_{ij}$. To implement the graphical and numerical assessment methods, $I(x_{ij} \le x)$ is replaced with $\frac{w_i(x)}{\sum_{i=1}^{n} w_i(x)}$ in the formulas for $W_j(x)$ and $\hat{W}_j(x)$.

You can perform the model checking described earlier for marginal models for dependent responses fit by generalized estimating equations (GEEs). Let $y_{ik}$ denote the $k$th measurement on the $i$th cluster, $i = 1, \ldots, K$, $k = 1, \ldots, n_i$, and let $\mathbf{x}_{ik}$ denote the corresponding vector of covariates. The marginal mean of the response $\mu_{ik} = \mathrm{E}(y_{ik})$ is assumed to depend on the covariate vector by

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

where $g$ is the link function.

Define the vector of residuals for the $i$th cluster as

$$\mathbf{e}_i = (e_{i1}, \ldots, e_{in_i})' = (y_{i1} - \hat{\mu}_{i1}, \ldots, y_{in_i} - \hat{\mu}_{in_i})'$$

You use the following extension of $W_j(x)$ defined earlier to check the functional form of the $j$th covariate:

$$W_j(x) = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \sum_{k=1}^{n_i} I(x_{ikj} \le x) e_{ik}$$

where $x_{ikj}$ is the $j$th component of $\mathbf{x}_{ik}$.

The null distribution of $W_j(x)$ can be approximated by the conditional distribution of

$$\hat{W}_j(x) = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \left\{ \sum_{k=1}^{n_i} I(x_{ikj} \le x) e_{ik} + \boldsymbol{\eta}'(x, \hat{\boldsymbol{\beta}}) \mathbf{I}_0^{-1} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \mathbf{e}_i \right\} Z_i$$

where $\hat{\mathbf{D}}_i$ and $\hat{\mathbf{V}}_i$ are defined as in the section “Generalized Estimating Equations” on page 3663 with the unknown parameters replaced by their estimated values,

$$\boldsymbol{\eta}(x, \boldsymbol{\beta}) = -\sum_{i=1}^{K}\sum_{k=1}^{n_i} I(x_{ikj} \leq x)\frac{\partial \mu_{ik}}{\partial \boldsymbol{\beta}}$$

$$\mathbf{I}_0 = \sum_{i=1}^{K} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i$$

and $Z_i, i = 1, \ldots, K$, are independent $N(0, 1)$ random variables. You replace $x_{ikj}$ with the linear predictor $\mathbf{x}_{ik}' \hat{\boldsymbol{\beta}}$ in the preceding formulas to check the link function.

## Case Deletion Diagnostic Statistics

For ordinary generalized linear models, regression diagnostic statistics developed by Williams (1987) can be requested in an output data set or in the OBSTATS table by specifying the DIAGNOSTICS | INFLUENCE option in the MODEL statement. These diagnostics measure the influence of an individual observation on model fit, and generalize the one-step diagnostics developed by Pregibon (1981) for the logistic regression model for binary data.

Preisser and Qaqish (1996) further generalized regression diagnostics to apply to models for correlated data fit by generalized estimating equations (GEEs), where the influence of entire clusters of correlated observations, or the influence of individual observations within a cluster, is measured. These diagnostic statistics can be requested in an output data set or in the OBSTATS table if a model for correlated data is specified with a REPEATED statement.

The next two sections use the following notation:

$\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of the regression parameters $\boldsymbol{\beta}$, or, in the case of correlated data, the solution of the GEEs.

$\hat{\boldsymbol{\beta}}_{[i]}$ is the corresponding estimate evaluated with the *i*th observation deleted, or, in the case of correlated data, with the *i*th cluster deleted.

$p$ is the dimension of the regression parameter vector $\boldsymbol{\beta}$.

$r_{pi}$ is the standardized Pearson residual $\frac{y_i - \mu_i}{\sqrt{v_i(1-h_i)}}$, where $v_i$ is the variance of the *i*th response and $h_i$ is the leverage defined in the section "H | LEVERAGE" on page 3676.

$v_i$ is the variance of response $i$, $\text{Var}(Y_i) = \phi V(\mu_i)$, where $V(\mu)$ is the variance function and $\phi$ is the dispersion parameter.

$w_i$ is the prior weight of the *i*th observation specified with the WEIGHT statement. If there is no WEIGHT statement, $w_i = 1$ for all $i$.

All unknown quantities are replaced by their estimated values in the following two sections.

## Diagnostics for Ordinary Generalized Linear Models

The following statistics are available for generalized linear models.

### DFBETA

The DFBETA statistic for measuring the influence of the $i$th observation is defined as the one-step approximation to the difference in the MLE of the regression parameter vector and the MLE of the regression parameter vector without the $i$th observation. This one-step approximation assumes a Fisher scoring step, and is given by

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]} \approx \mathrm{DFBETA}_i = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_i'\mathbf{W}_i^{\frac{1}{2}}(1-h_i)^{-\frac{1}{2}} r_{pi}$$

where $h_i$ is the leverage defined in the section "H | LEVERAGE" on page 3676.

### DFBETAS

The standardized DFBETA statistic for assessing the influence of the $i$th observation on the $j$th regression parameter is defined as the DFBETA statistic for the $j$th parameter divided by its estimated standard deviation, where the standard deviation is estimated from all the data.

$$\mathrm{DFBETAS}_{ij} = \mathrm{DFBETA}_{ij}/\hat{\sigma}(\beta_j)$$

### DOBS | COOKD | COOKSD

In normal linear regression, the influence of observation $i$ can be measured by Cook's distance (Cook and Weisberg 1982). A measure of influence of observation $i$ for generalized linear models that is equivalent to Cook's distance for normal linear regression is given by

$$\mathrm{DOBS}_i = p^{-1} h_i (1-h_i)^{-1} r_{pi}^2$$

where $h_i$ is the leverage defined in the section "H | LEVERAGE" on page 3676. This measure is the one-step approximation to $2p^{-1}[L(\hat{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}}_{[i]})]$, where $L(\boldsymbol{\beta})$ is the log likelihood evaluated at $\boldsymbol{\beta}$.

### H | LEVERAGE

The Fisher scores, or expected, weight for observation $i$ is $w_{ei} = \frac{w_i}{\phi V(\mu_i)(g'(\mu_i))^2}$. Let $\mathbf{W}$ be the diagonal matrix with $w_{ei}$ as the $i$th diagonal. The leverage $h_i$ of the $i$th observation is defined as the $i$th diagonal element of the hat matrix

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

## Diagnostics for Models Fit by Generalized Estimating Equations (GEEs)

The diagnostic statistics in this section were developed by Preisser and Qaqish (1996). See the section "Generalized Estimating Equations" on page 3663 for further information and notation for generalized estimating equations (GEEs). The following additional notation is used in this section.

Partition the design matrix $\mathbf{X}$ and response vector $\mathbf{Y}$ by cluster; that is, let $\mathbf{X} = (X_1', \ldots, X_K')'$, and $\mathbf{Y} = (Y_1', \ldots, Y_K')'$ corresponding to the $K$ clusters.

Let $n_i$ be the number of responses for cluster *i*, and denote by $N = \sum_{i=1}^{K} n_i$ the total number of observations. Denote by $A_i$ the $n_i \times n_i$ diagonal matrix with $V(\mu_{ij})$ as the *j*th diagonal element. If there is a WEIGHT statement, the diagonal element of $A_i$ is $V(\mu_{ij})/w_{ij}$, where $w_{ij}$ is the specified weight of the *j*th observation in the *i*th cluster. Let $\mathbf{B}$ the $N \times N$ diagonal matrix with $g'(\mu_{ij})$ as diagonal elements, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$. Let $\mathbf{B}_i$ the $n_i \times n_i$ diagonal matrix corresponding to cluster *i* with $g'(\mu_{ij})$ as the *j*th diagonal element.

Let $\mathbf{W}$ be the $N \times N$ block diagonal weight matrix whose *i*th block, corresponding to the *i*th cluster, is the $n_i \times n_i$ matrix

$$\mathbf{W}_{ei} = \mathbf{B}_i^{-1}\mathbf{A}_i^{-\frac{1}{2}}\mathbf{R}_i^{-1}(\hat{\boldsymbol{\alpha}})\mathbf{A}_i^{-\frac{1}{2}}\mathbf{B}_i^{-1}$$

where $\mathbf{R}_i$ is the working correlation matrix for cluster *i*.

Let

$$Q_i = \mathbf{X}_i(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_i'$$

where $\mathbf{X}_i$ is the $n_i \times p$ design matrix corresponding to cluster *i*.

Define the adjusted residual vector as

$$\mathbf{E} = \mathbf{B}(\mathbf{Y} - \hat{\boldsymbol{\mu}})$$

and $\mathbf{E}_i = \mathbf{B}_i(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)$, the estimated residual for the *i*th cluster.

Let the subscript $[i]$ denote estimates evaluated without the *i*th cluster, $[it]$ estimates evaluated using all the data except the *t*th observation of the *i*th cluster, and let $i[t]$ denote matrices corresponding to the *i*th cluster without the *t*th observation.

The following statistics are available for generalized estimating equation models.

### CH | CLUSTERH | CLEVERAGE

The leverage of cluster *i* is contained in the matrix $\mathbf{H}_i = \mathbf{Q}_i\mathbf{W}_{ei}$, and is summarized by the trace of $\mathbf{H}_i$,

$$ch_i = \mathrm{tr}(\mathbf{H}_i)$$

The leverage $h_i$ of the *t*th observation in the *i*th cluster is the *t*th diagonal element of $\mathbf{H}_i$.

### DFBETAC

The effect of deleting cluster *i* on the estimated parameter vector is given by the following one-step approximation for $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]}$:

$$\mathrm{DBETAC}_i = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_i'(\mathbf{W}_{ei}^{-1} - \mathbf{Q}_i)^{-1}\mathbf{E}_i$$

### DFBETACS

The cluster deletion statistic DFBETAC can be standardized using the variances of $\hat{\boldsymbol{\beta}}$ based on the complete data. The standardized one-step approximation for the change in $\hat{\beta}_j$ due to deletion of cluster *i* is

$$\mathrm{DBETACS}_{ij} = \frac{\mathrm{DBETAC}_{ij}}{\hat{\phi}[(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}]_{jj}^{\frac{1}{2}}}$$

### DFBETA

Partition the matrices $\mathbf{W}_{ei}$ and $\mathbf{V}_i$ as

$$\mathbf{W}_{ei} = \begin{pmatrix} W_{eit} & \mathbf{W}_{eit[t]} \\ \mathbf{W}_{ei[t]t} & \mathbf{W}_{ei[t]} \end{pmatrix}$$

$$\mathbf{V}_i = \mathbf{W}_{ei}^{-1} = \begin{pmatrix} \mathbf{V}_{it} & \mathbf{V}_{it[t]} \\ \mathbf{V}_{i[t]t} & \mathbf{V}_{i[t]} \end{pmatrix}$$

and let $\mathbf{E}_{it} = \mathbf{B}_{it}(\mathbf{Y}_{it} - \hat{\mu}_{it})$ and $\mathbf{E}_{i[t]} = \mathbf{B}_{i[t]}(\mathbf{Y}_{i[t]} - \hat{\mu}_{i[t]})$.

The effect of deleting the $t$th observation from the $i$th cluster is given by the following one-step approximation to $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[it]}$:

$$\mathbf{DBETAO}_{it} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\tilde{\mathbf{X}}'_{it}\frac{\tilde{E}_{it}}{W_{eit}^{-1} - \tilde{Q}_{it}}$$

where $\tilde{\mathbf{X}}_{it} = \mathbf{X}_{it} - \mathbf{V}_{it[t]}\mathbf{V}_{i[t]}^{-1}\mathbf{X}_{i[t]}$, $\tilde{Q}_{it} = \tilde{\mathbf{X}}_{it}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\tilde{X}'_{it}$, and $\tilde{E}_{it} = \mathbf{E}_{it} - \mathbf{V}_{it[t]}\mathbf{V}_{i[t]}^{-1}\mathbf{E}_{i[t]}$. Note that $W_{eit}$, $\tilde{Q}_{it}$, and $\tilde{E}_{it}$ are scalars.

### DFBETAS

The observation deletion statistic DFBETA can be standardized using the variances of $\hat{\boldsymbol{\beta}}$ based on the complete data. The standardized one-step approximation for the change in $\hat{\beta}_j$ due to deletion of observation $t$ in cluster $i$ is

$$\text{DBETAOS}_{itj} = \frac{\text{DBETAO}_{itj}}{\hat{\phi}[(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}]_{jj}^{\frac{1}{2}}}$$

### DCLS | CLUSTERCOOKD | CLUSTERCOOKSD

A measure of the standardized influence of the subset $m$ of observations on the overall fit is $(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[m]})'(\mathbf{X}'\mathbf{W}\mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[m]})/p\hat{\phi}$. For deletion of cluster $i$, this is approximated by

$$\text{DCLS}_i = \mathbf{E}'_i(\mathbf{W}_{ei}^{-1} - \mathbf{Q}_i)^{-1})\mathbf{Q}_i(\mathbf{W}_{ei}^{-1} - \mathbf{Q}_i)^{-1})\mathbf{E}_i/p\hat{\phi}$$

### DOBS | COOKD | COOKSD

The measure of overall fit in the section "DCLS | CLUSTERCOOKD | CLUSTERCOOKSD" on page 3678 for the deletion of the $t$th observation in the $i$th cluster is approximated by

$$\text{DOBS}_{it} = \frac{\tilde{E}_{it}^2\tilde{Q}_{it}}{p\hat{\phi}(W_{eit}^{-1} - \tilde{Q}_{it})^2}$$

where $\tilde{E}_{it}$, $\tilde{Q}_{it}$, and $W_{eit}$ are defined in the section "DFBETA" on page 3678. In the case of the independence working correlation, this is equal to the measure for ordinary generalized linear models defined in the section "DOBS | COOKD | COOKSD" on page 3676.

### MCLS | CLUSTERDFIT

A studentized distance measure of the type defined in the section "DCLS | CLUSTERCOOKD | CLUSTERCOOKSD" on page 3678 of the influence of the *i*th cluster is given by

$$\mathrm{MCLS}_i = \mathbf{E}_i'(\mathbf{W}_{ei}^{-1} - \mathbf{Q}_i)^{-1}\mathbf{H}_i\mathbf{E}_i / p\hat{\phi}$$

## Bayesian Analysis

In generalized linear models, the response has a probability distribution from a family of distributions 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. The canonical parameters $\theta$ depend only on the means of the response $\mu_i$, which are related to the regression parameters $\beta$ through the link function $g(\mu_i) = x'\beta$. The additional parameter $\phi$ is the dispersion parameter. The GENMOD procedure estimates the regression parameters and the scale parameter $\sigma = \phi^{\frac{1}{2}}$ by maximum likelihood. However, the GENMOD procedure can also provide Bayesian estimates of the regression parameters and either the scale $\sigma$, the dispersion $\phi$, or the precision $\tau = \phi^{-1}$ by sampling from the posterior distribution. Except where noted, the following discussion applies to either $\sigma$, $\phi$, or $\tau$, although $\phi$ is used to illustrate the formulas. Note that the Poisson and binomial distributions do not have a dispersion parameter, and the dispersion is considered to be fixed at $\phi = 1$. The ASSESS, CONTRAST, ESTIMATE, OUTPUT, and REPEATED statements, if specified, are ignored. Also ignored are the PLOTS= option in the PROC GENMOD statement and the following options in the MODEL statement: ALPHA=, CORRB, COVB, TYPE1, TYPE3, SCALE=DEVIANCE (DSCALE), SCALE=PEARSON (PSCALE), OBSTATS, RESIDUALS, XVARS, PREDICTED, DIAGNOSTICS, and SCALE= for Poisson and binomial distributions. The multinomial and zero-inflated Poisson distributions are not available for Bayesian analysis.

See the section "Assessing Markov Chain Convergence" on page 162 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for information about assessing the convergence of the chain of posterior samples.

Several algorithms, specified with the SAMPLING= option in the BAYES statement, are available in GENMOD for drawing samples from the posterior distribution.

### ARMS Algorithm for Gibbs Sampling

This section provides details for Bayesian analysis by Gibbs sampling in generalized linear models. See the section "Gibbs Sampler" on page 157 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for a general discussion of Gibbs sampling. See Gilks, Richardson, and Spiegelhalter (1996) for a discussion of applications of Gibbs sampling to a number of different models, including generalized linear models.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ be the parameter vector. For generalized linear models, the $\theta_i$s are the regression coefficients $\beta_i$s and the dispersion parameter $\phi$. Let $L(D|\boldsymbol{\theta})$ be the likelihood function, where $D$ is the observed data. Let $\pi(\boldsymbol{\theta})$ be the prior distribution. The full conditional distribution of $[\theta_i|\theta_j, i \neq j]$ is proportional to the joint distribution; that is,

$$\pi(\theta_i|\theta_j, i \neq j, D) \propto L(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

For instance, the one-dimensional conditional distribution of $\theta_1$ given $\theta_j = \theta_j^*, 2 \leq j \leq k$, is computed as

$$\pi(\theta_1|\theta_j = \theta_j^*, 2 \leq j \leq k, D) = L(D|(\boldsymbol{\theta} = (\theta_1, \theta_2^*, \ldots, \theta_k^*)')p(\boldsymbol{\theta} = (\theta_1, \theta_2^*, \ldots, \theta_k^*)')$$

Suppose you have a set of arbitrary starting values $\{\theta_1^{(0)}, \ldots, \theta_k^{(0)}\}$. Using the ARMS (adaptive rejection Metropolis sampling) algorithm (Gilks and Wild 1992; Gilks, Best, and Tan 1995), you can do the following:

$$\text{draw } \theta_1^{(1)} \text{ from } [\theta_1|\theta_2^{(0)}, \ldots, \theta_k^{(0)}]$$

$$\text{draw } \theta_2^{(1)} \text{ from } [\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}]$$

$$\ldots$$

$$\text{draw } \theta_k^{(1)} \text{ from } [\theta_k|\theta_1^{(1)}, \ldots, \theta_{k-1}^{(1)}]$$

This completes one iteration of the Gibbs sampler. After one iteration, you have $\{\theta_1^{(1)}, \ldots, \theta_k^{(1)}\}$. After $n$ iterations, you have $\{\theta_1^{(n)}, \ldots, \theta_k^{(n)}\}$. PROC GENMOD implements the ARMS algorithm provided by Gilks (2003) to draw a sample from a full conditional distribution. See the section "Adaptive Rejection Sampling Algorithm" on page 158 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for more information about the ARMS algorithm.

## Gamerman Algorithm

The Gamerman algorithm, unlike a Gibbs sampling algorithm, samples parameters from their multivariate posterior conditional distribution. The algorithm uses the structure of generalized linear models to efficiently sample from the posterior distribution of the model parameters. For a detailed description and explanation of the algorithm, see Gamerman (1997) and the section "Gamerman Algorithm" on page 160 in Chapter 8, "Introduction to Bayesian Analysis Procedures." The Gamerman algorithm is the default method used to sample from the posterior distribution. See any of the introductory references in Chapter 8, "Introduction to Bayesian Analysis Procedures," for a discussion of conjugate prior distributions for a linear model with the normal distribution.

## Independence Metropolis Algorithm

The independence Metropolis algorithm is another sampling algorithm that draws multivariate samples from the posterior distribution. See the section "Independence Sampler" on page 160 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for more details.

## Posterior Samples Output Data Set

You can output posterior samples into a SAS data set through ODS. The following SAS statement outputs the posterior samples into the SAS data set Post:

**ODS OUTPUT POSTERIORSAMPLE=***Post*

You can alternatively create the SAS data set Post with the **OUTPOST=***Post* option in the BAYES statement.

The data set also includes the variables LogPost and LogLike, which represent the log of the posterior likelihood and the log of the likelihood, respectively.

## Priors for Model Parameters

The model parameters are the regression coefficients and the dispersion parameter (or the precision or scale), if the model has one. The priors for the dispersion parameter and the priors for the regression coefficients are assumed to be independent, while you can have a joint multivariate normal prior for the regression coefficients.

### *Dispersion, Precision, or Scale Parameter*

**Gamma Prior** The gamma distribution $G(a, b)$ has a probability density function

$$f(u) = \frac{b(bu)^{a-1}\mathrm{e}^{-bu}}{\Gamma(a)}, \qquad u > 0$$

where $a$ is the shape parameter and $b$ is the inverse-scale parameter. The mean is $\frac{a}{b}$ and the variance is $\frac{a}{b^2}$.

**Improper Prior** The joint prior density is given by

$$p(u) \propto u^{-1}, \qquad u > 0$$

**Inverse Gamma Prior** The inverse gamma distribution $\mathrm{IG}(a, b)$ has a probability density function

$$f(u) = \frac{b^a}{\Gamma(a)} u^{-(a+1)}\mathrm{e}^{-b/u}, \qquad u > 0$$

where $a$ is the shape parameter and $b$ is the scale parameter. The mean is $\frac{b}{a-1}$ if $a > 1$, and the variance is $\frac{b^2}{(a-1)^2(a-2)}$ if $a > 2$.

### *Regression Coefficients*

Let $\boldsymbol{\beta}$ be the regression coefficients.

**Jeffreys' Prior** The joint prior density is given by

$$p(\boldsymbol{\beta}) \propto |\mathbf{I}(\boldsymbol{\beta})|^{\frac{1}{2}}$$

where $\mathbf{I}(\boldsymbol{\beta})$ is the Fisher information matrix for the model. If the underlying model has a scale parameter (for example, a normal linear regression model), then the Fisher information matrix is computed with the scale parameter set to a fixed value of one.

If you specify the CONDITIONAL option, then Jeffreys' prior, conditional on the current Markov chain value of the generalized linear model precision parameter $\tau$, is given by

$$|\tau\mathbf{I}(\boldsymbol{\beta})|^{\frac{1}{2}}$$

where $\tau$ is the model precision parameter.

See Ibrahim and Laud (1991) for a full discussion, with examples, of Jeffreys' prior for generalized linear models.

**Normal Prior** Assume $\boldsymbol{\beta}$ has a multivariate normal prior with mean vector $\boldsymbol{\beta}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. The joint prior density is given by

$$p(\boldsymbol{\beta}) \propto \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}$$

If you specify the CONDITIONAL option, then, conditional on the current Markov chain value of the generalized linear model precision parameter $\tau$, the joint prior density is given by

$$p(\boldsymbol{\beta}) \propto \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\tau\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}$$

**Uniform Prior** The joint prior density is given by

$$p(\boldsymbol{\beta}) \propto 1$$

## Deviance Information Criterion

Let $\theta_i$ be the model parameters at iteration $i$ of the Gibbs sampler and let $\mathrm{LL}(\theta_i)$ be the corresponding model log likelihood. PROC GENMOD computes the following fit statistics defined by Spiegelhalter et al. (2002):

- Effective number of parameters:

$$p_D = \overline{\mathrm{LL}(\theta)} - \mathrm{LL}(\bar{\theta})$$

- Deviance information criterion (DIC):

$$\mathrm{DIC} = \overline{\mathrm{LL}(\theta)} + p_D$$

where

$$\begin{aligned} \overline{\mathrm{LL}(\theta)} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathrm{LL}(\theta_i) \\ \bar{\theta} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \theta_i \end{aligned}$$

PROC GENMOD uses the full log likelihoods defined in the section “Log-Likelihood Functions” on page 3644, with all terms included, for computing the DIC.

## Posterior Distribution

Denote the observed data by $D$.

The posterior distribution is

$$\pi(\boldsymbol{\beta}|D) \propto L_P(D|\boldsymbol{\beta})p(\boldsymbol{\beta})$$

where $L_P(D|\boldsymbol{\beta})$ is the likelihood function with regression coefficients $\boldsymbol{\beta}$ as parameters.

## Starting Values of the Markov Chains

When the BAYES statement is specified, PROC GENMOD generates one Markov chain containing the approximate posterior samples of the model parameters. Additional chains are produced when the Gelman-Rubin diagnostics are requested. Starting values (or initial values) can be specified in the INITIAL= data set in the BAYES statement. If INITIAL= option is not specified, PROC GENMOD picks its own initial values for the chains.

Denote $[x]$ as the integral value of x. Denote $\hat{s}(X)$ as the estimated standard error of the estimator $X$.

### *Regression Coefficients*

For the first chain that the summary statistics and regression diagnostics are based on, the default initial values are estimates of the mode of the posterior distribution. If the INITIALMLE option is specified, the initial values are the maximum likelihood estimates; that is,

$$\beta_i^{(0)} = \hat{\beta}_i$$

Initial values for the $r$th chain ($r \geq 2$) are given by

$$\beta_i^{(0)} = \hat{\beta}_i \pm \left(2 + \left[\frac{r}{2}\right]\right)\hat{s}(\hat{\beta}_i)$$

with the plus sign for odd $r$ and minus sign for even $r$.

### *Dispersion, Scale, or Precision Parameter $\lambda$*

Let $\lambda$ be the generalized linear model parameter you choose to sample, either the dispersion, scale, or precision parameter. Note that the Poisson and binomial distributions do not have this additional parameter.

For the first chain that the summary statistics and regression diagnostics are based on, the default initial values are estimates of the mode of the posterior distribution. If the INITIALMLE option is specified, the initial values are the maximum likelihood estimates; that is,

$$\lambda^{(0)} = \hat{\lambda}$$

The initial values of the $r$th chain ($r \geq 2$) are given by

$$\lambda^{(0)} = \hat{\lambda}\mathrm{e}^{\pm\left([\frac{r}{2}]+2\right)\hat{s}(\hat{\lambda})}$$

with the plus sign for odd $r$ and minus sign for even $r$.

## OUTPOST= Output Data Set

The OUTPOST= data set contains the generated posterior samples. There are 3+$n$ variables, where $n$ is the number of model parameters. The variable Iteration represents the iteration number, the variable LogLike contains the log of the likelihood, and the variable LogPost contains the log of the posterior. The other $n$ variables represent the draws of the Markov chain for the model parameters.

## Exact Logistic and Exact Poisson Regression

The theory of exact logistic regression, also called exact conditional logistic regression, is described in the section "Exact Conditional Logistic Regression" on page 6039 in Chapter 79, "The LOGISTIC Procedure." The following discussion of exact Poisson regression, also called exact conditional Poisson regression, uses the notation given in that section.

Note that in exact logistic regression, the coefficients $C(\boldsymbol{t})$ are the number of possible response vectors $\boldsymbol{y}$ that generate $\boldsymbol{t}$: $C(\boldsymbol{t}) = ||\{\boldsymbol{y} : \boldsymbol{y}'\mathbf{X} = \boldsymbol{t}'\}||$. However, when performing an exact Poisson regression, this value is replaced by

$$C(\boldsymbol{t}) = \sum_{\Omega} \prod_{i=1}^{n} \frac{N_i^{y_i}}{y_i!}$$

where $\Omega = \{\boldsymbol{y} \colon \boldsymbol{y}'\mathbf{X} = \boldsymbol{t}\}$ and $N_i = \exp(o_i)$ is the exponential of the offset $o_i$ for observation $i$. If an offset variable is not specified, then $N_i = 1$.

The probability density function (PDF) for $\mathbf{T}$ is created by summing over all candidate sequences $\boldsymbol{y}$ that generate an observable $\boldsymbol{t}$

$$\Pr(\mathbf{T} = \boldsymbol{t}) = \frac{C(\boldsymbol{t}) \exp(\boldsymbol{t}'\boldsymbol{\beta})}{\prod_{i=1}^{n} \exp(N_i e^{\boldsymbol{x}_i'\boldsymbol{\beta}})}$$

However, the conditional likelihood of $\mathbf{T}_\mathrm{I}$ given $\mathbf{T}_\mathrm{N} = \boldsymbol{t}_\mathrm{N}$ has the same form as that for exact logistic regression.

For details about hypothesis testing and estimation, see the sections "Hypothesis Tests" on page 6041 and "Inference for a Single Parameter" on page 6042 in Chapter 79, "The LOGISTIC Procedure." See the section "Computational Resources for Exact Logistic Regression" on page 6050 in Chapter 79, "The LOGISTIC Procedure," for some computational notes about exact analyses.

In exact logistic binary regression, each component $y_i, i = 1, \ldots, n$, of $\boldsymbol{y}$ can take a value of 0 or 1, so there are a finite number, $2^n$, of candidate $\boldsymbol{y}$ vectors to be considered. Since a Poisson-distributed response variable can take an infinite number of values, exact Poisson regression should evaluate an infinite number of $\boldsymbol{y}$ vectors. However, by identifying the maximum value of $y_i$ to check, $S_i$, for each observation $i$, the number of candidate $\boldsymbol{y}$ vectors to check is reduced to $\prod_{i=1}^{n} S_i$. On a practical level, as $S_i$ becomes large the probability of the Poisson random variable achieving this value drops to zero, so $S_i$ can be thought of as the point at which the value does not matter. You can provide these maxima by specifying either an OFFSET= variable, $o_i$, or an EXACTMAX= variable, $e_i$, or you can let the algorithm choose a maximum for you. The way these two options interact to provide a maximum is described in the following list:

1. If an EXACTMAX= variable is specified, then $S_i = e_i$.
2. If the EXACTMAX option is specified without a variable, or if neither the EXACTMAX= nor OFFSET= options are specified, then you must also condition out the intercept or you must specify the STRATA statement. If you are conditioning out the intercept, then every $S_i$ has an effective maximum of $\sum_{i=1}^{n} f_i y_{0i}$, where $\boldsymbol{y}_0$ is the observed response and $f_i$ is the frequency of the observation; this is the sufficient statistic for the intercept term. If you are performing a stratified analysis, these sums are computed within each stratum.

3. If an offset variable is specified and the EXACTMAX option is not specified, then $S_i = \max(\text{ceil}(\exp(o_i)), 1)$. For example, if you have $N_i$ rats in cage $i$ and you are modeling the proportion that acquire a disease, then you would set your offsets to $o_i = \log(N_i)$ so that $S_i = N_i$. In this case, the offsets must also satisfy $S_i \geq y_i$.

## OUTDIST= Output Data Set

The OUTDIST= data set contains every exact conditional distribution necessary to process the corresponding EXACT statement. For example, the following statements create one distribution for the x1 parameter and another for the x2 parameters, and produce the data set dist shown in Table 51.12:

```
data test;
   input y x1 x2 count;
   datalines;
0 0 0 1
1 0 0 1
0 1 1 2
1 1 1 1
1 0 2 3
1 1 2 1
1 2 0 3
1 2 1 2
1 2 2 1
;

proc genmod data=test exactonly;
   class x2 / param=ref;
   model y=x1 x2 / d=b;
   exact x1 x2/ outdist=dist;
run;
proc print data=dist;
run;
```

**Table 51.12** OUTDIST= Data Set

| Obs | x1 | x20 | x21 | Count | Score | Prob |
|---|---|---|---|---|---|---|
| 1 | . | 0 | 0 | 3 | 5.81151 | 0.03333 |
| 2 | . | 0 | 1 | 15 | 1.66031 | 0.16667 |
| 3 | . | 0 | 2 | 9 | 3.12728 | 0.10000 |
| 4 | . | 1 | 0 | 15 | 1.46523 | 0.16667 |
| 5 | . | 1 | 1 | 18 | 0.21675 | 0.20000 |
| 6 | . | 1 | 2 | 6 | 4.58644 | 0.06667 |
| 7 | . | 2 | 0 | 19 | 1.61869 | 0.21111 |
| 8 | . | 2 | 1 | 2 | 3.27293 | 0.02222 |
| 9 | . | 3 | 0 | 3 | 6.27189 | 0.03333 |
| 10 | 2 | . | . | 6 | 3.03030 | 0.12000 |
| 11 | 3 | . | . | 12 | 0.75758 | 0.24000 |
| 12 | 4 | . | . | 11 | 0.00000 | 0.22000 |
| 13 | 5 | . | . | 18 | 0.75758 | 0.36000 |
| 14 | 6 | . | . | 3 | 3.03030 | 0.06000 |

The first nine observations in the dist data set contain an exact distribution for the parameters of the x2 effect (hence the values for the x1 parameter are missing), and the remaining five observations are for the x1 parameter. If a joint distribution was created, there would be observations with values for both the x1 and x2 parameters. For CLASS variables, the corresponding parameters in the dist data set are identified by concatenating the variable name with the appropriate classification level.

The data set contains the possible sufficient statistics of the parameters for the effects specified in the EXACT statement, and the Count variable contains the number of different responses that yield these statistics. In particular, there are six possible response vectors $\mathbf{y}$ for which the dot product $\mathbf{y}'\mathbf{x1}$ was equal to 2, and for which $\mathbf{y}'\mathbf{x20}$, $\mathbf{y}'\mathbf{x21}$, and $\mathbf{y}'\mathbf{1}$ were equal to their actual observed values (displayed in the “Sufficient Statistics” table).

**NOTE:** If you are performing an exact Poisson analysis, then the Count variable is replaced by a variable named Weight.

When hypothesis tests are performed on the parameters, the Prob variable contains the probability of obtaining that statistic (which is just the count divided by the total count), and the Score variable contains the score for that statistic.

The OUTDIST= data set can contain a different exact conditional distribution for each specified EXACT statement. For example, consider the following EXACT statements:

```
exact 'O1'   x1    /           outdist=o1;
exact 'OJ12' x1 x2 / jointonly outdist=oj12;
exact 'OA12' x1 x2 / joint     outdist=oa12;
exact 'OE12' x1 x2 / estimate  outdist=oe12;
```

The O1 statement outputs a single exact conditional distribution. The OJ12 statement outputs only the joint distribution for x1 and x2. The OA12 statement outputs three conditional distributions: one for x1, one for x2, and one jointly for x1 and x2. The OE12 statement outputs two conditional distributions: one for x1 and the other for x2. Data set oe12 contains both the x1 and x2 variables; the distribution for x1 has missing values in the x2 column while the distribution for x2 has missing values in the x1 column.

## Response Level Ordering

For binary response data, PROC GENMOD fits the following model by default:

$$\Phi^{-1}\left(\frac{p-C}{1-C}\right) = \mathrm{x}'\boldsymbol{\beta}$$

where $p$ is the probability of the response level identified as the first level in the "Response Profile" table in the output and $\Phi$ is the normal cumulative distribution function. By default, the covariate vector $\mathbf{x}$ contains an intercept term. This is sometimes called Abbot's formula.

Because of the symmetry of the normal (and logistic) distribution, the effect of reversing the order of the two response values is to change the signs of $\boldsymbol{\beta}$ in the preceding equation.

By default, response levels appear in ascending, sorted order (that is, the lowest level appears first, and then the next lowest, and so on). There are a number of ways that you can control the sort order of the response categories and, therefore, which level is assigned the first ordered level. One of the most common sets of response levels is $\{0,1\}$, with 1 representing the event with the probability that is to be modeled.

Consider the example where Y takes the values 1 and 0 for event and nonevent, respectively, and EXPOSURE is the explanatory variable. By default, PROC GENMOD assigns the first ordered level to response level 0, causing the probability of the nonevent to be modeled. There are several ways to change this.

Besides recoding the variable Y, you can do the following:

- Explicitly state which response level is to be modeled by using the response variable option EVENT= in the MODEL statement:

```
      model Y(event='1') = Exposure;
```

- Specify the nonevent category for the response variable in the response variable option REF= in the MODEL statement:

```
      model Y(ref='0') = Exposure;
```

- Specify the response variable option DESCENDING in the MODEL statement to assign the lowest ordered value to Y=1:

```
      model Y(descending)=Exposure;
```

- Assign a format to Y such that the first formatted value (when the formatted values are put in sorted order) corresponds to the event. For the following example, Y=0 could be assigned formatted value 'nonevent' and Y=1 could be assigned formatted value 'event.' Since ORDER=FORMATTED by default, Y=1 becomes the first ordered level.

```
proc format;
   value disease 1='event' 0='nonevent';
run;
proc genmod;
   model y=exposure / dist=bin;
   format y disease.;
run;
```

- Arrange the input data set so that Y=1 appears first and use the ORDER=DATA option in the PROC GENMOD statement. Because ORDER=DATA sorts levels in order of their appearance in the data set, Y=1 becomes the first ordered level. Note that this option causes classification variables to be sorted by their order of appearance in the data set, also.

## Missing Values

For generalized linear models, PROC GENMOD ignores any observation with a missing value for any variable involved in the model. You can score an observation in an output data set by setting only the response value to missing. For models fit with generalized estimating equations (GEEs), observations with missing values within a cluster are not used, and all available pairs are used in estimating the working correlation matrix. Clusters with fewer observations than the full cluster size are treated as having missing observations occurring at the end of the cluster. You can specify the order of missing observations with the WITHINSUBJECT= option. See the section "Missing Data" on page 3666 for more information about missing values in GEEs.

## Displayed Output for Classical Analysis

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 3699.

### Model Information

The "Model Information" table displays the two-level data set name, the response distribution, the link function, the response variable name, the offset variable name, the frequency variable name, the scale weight variable name, the number of observations used, the number of events if events/trials format is used for response, the number of trials if events/trials format is used for response, the sum of frequency weights, the number of missing values in data set, and the 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: the iteration number, the ridge value, the log likelihood, and values of all parameters in the model.

## Criteria for Assessing Goodness of Fit

In the “Criteria for Assessing Goodness of Fit” table, PROC GENMOD displays the degrees of freedom for deviance and Pearson’s chi-square, equal to the number of observations minus the number of regression parameters estimated, the deviance, the deviance divided by degrees of freedom, the scaled deviance, the scaled deviance divided by degrees of freedom, Pearson’s chi-square, Pearson’s chi-square divided by degrees of freedom, the scaled Pearson’s chi-square, the scaled Pearson’s chi-square divided by degrees of freedom, the log likelihood (excludes factorial terms) the full log likelihood, the Akaike information criterion, the corrected Akaike information criterion, and the Bayesian information criterion. The information in this table is valid only for maximum likelihood model fitting, and the table is not printed if the REPEATED statement is specified.

## 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, and are displayed only if the PRINTMLE option in the REPEATED statement is specified. For each parameter in the model, PROC GENMOD displays the parameter name, as follows:

- 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

In addition, PROC GENMOD displays the degrees of freedom for the parameter, the estimate value, the standard error, the Wald chi-square value, the *p*-value based on the chi-square distribution, and the confidence limits (Wald or profile likelihood) for parameters.

### Lagrange Multiplier Statistics

If you specify that either the model intercept or the scale parameter is fixed, for those distributions that have a distribution scale parameter, the GENMOD procedure displays a table of Lagrange multiplier, or score, statistics for testing the validity of the constrained parameter that contains the test statistic, and the *p*-value.

### 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 parameter identification number, the iteration number, the 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 that summarizes profile likelihood-based confidence intervals for all parameters. For each parameter in the model, the table displays the confidence coefficient, the parameter identification number, lower and upper endpoints of confidence intervals for the parameter, and values of all other parameters at the solution.

### LR Statistics for Type 1 Analysis

If you specify the TYPE1 model option, a table is displayed that contains the name of the effect, the deviance for the model including the effect and all previous effects, the degrees of freedom for the effect, the likelihood ratio statistic for testing the significance of the effect, and the *p*-value computed from the chi-square distribution with the effect's degrees of freedom.

If you specify either the SCALE=DEVIANCE or SCALE=PEARSON option in the MODEL statement, columns are displayed that contain the name of the effect, the deviance for the model including the effect and all previous effects, the numerator degrees of freedom, the denominator degrees of freedom, the chi-square statistic for testing the significance of the effect, the *p*-value computed from the chi-square distribution with numerator degrees of freedom, the *F* statistic for testing the significance of the effect, and the *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 that contains the effect name, the iteration number, the ridge value, the log likelihood, and values of all parameters.

### LR Statistics for Type 3 Analysis

If you specify the TYPE3 model option, a table is displayed that contains, for each effect in the model, the name of the effect, the likelihood ratio statistic for testing the significance of the effect, the degrees of freedom for the effect, and the *p*-value computed from the chi-square distribution.

If you specify either the SCALE=DEVIANCE or SCALE=PEARSON option in the MODEL statement, columns are displayed that contain the name of the effect, the likelihood ratio statistic for testing the significance of the effect, the *F* statistic for testing the significance of the effect, the numerator degrees of freedom, the denominator degrees of freedom, the *p*-value based on the *F* distribution, and the *p*-value computed from the chi-square distribution with the numerator's degrees of freedom.

### Wald Statistics for Type 3 Analysis

If you specify the TYPE3 and WALD model options, a table is displayed that contains the name of the effect, the degrees of freedom of the effect, the Wald statistic for testing the significance of the effect, and the *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 an index number that identifies the parameter, and 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. Residuals and case deletion diagnostic statistics are not available for the multinomial distribution. Case deletion diagnostics are not available for zero-inflated models.

For each observation in the input data set, the following are displayed:

- the value of the response variable
- the predicted value of the mean

- the value of the linear predictor The value of an OFFSET variable is added to the linear predictor.
- the estimated standard error of the linear predictor
- the value of the negative of the weight in the Hessian matrix at the final iteration. 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
- raw residual
- Pearson residual
- deviance residual
- standardized Pearson residual
- standardized deviance residual
- likelihood residual
- leverage
- Cook's distance statistic
- DFBETA statistic, for each parameter
- standardized DFBETA statistic, for each parameter
- zero-inflation probability for zero-inflated models
- response mean for zero-inflated models

## 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.

For each ESTIMATE statement, the table contains the contrast label, the estimated value of the contrast, the standard error of the estimate, the significance level $\alpha$, $(1-\alpha) \times 100\%$ confidence intervals for contrast, the Wald chi-square statistic for the contrast, and the $p$-value computed from the chi-square distribution. The mean of the contrast, defined as the inverse link function applied to the contrast, and $(1-\alpha) \times 100\%$ confidence intervals for the mean are also displayed.

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 contrast label, and 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 contrast label, the iteration number, the ridge value, the log likelihood, and 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.

A table is displayed that contains the contrast label, the degrees of freedom for the contrast, and the 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. Also displayed are the $p$-value computed from the chi-square distribution, and 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 are displayed that contain the contrast label, the likelihood ratio statistic for testing the significance of the contrast, the $F$ statistic for testing the significance of the contrast, the numerator degrees of freedom, the denominator degrees of freedom, the $p$-value based on the $F$ distribution, and the $p$-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, the "Coefficients for *effect* Least Squares Means" table is displayed, where *effect* is the effect specified in the LSMEANS statement. The table contains the effect names and the rows of least squares means coefficients.

### Least Squares Means

If you specify the LSMEANS statement, the "Least Squares Means" table is displayed. The table contains for each effect the following: the effect name, and for each level of each effect the following:

- 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 "GEE Model Information" table displays the correlation structure of the working correlation matrix or the log odds ratio structure, the within-subject effect, the subject effect, the number of clusters, the correlation matrix dimension, and the 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 the "Log Odds Ratio Parameter Information" 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, the "Iteration History For GEE Parameter Estimates" table is displayed. The table contains the parameter identification number, the iteration number, and 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 Evaluation Of The Generalized Gradient And Hessian" table.

### GEE Parameter Estimate Covariance Matrices

If you specify the REPEATED statement and the COVB option, PROC GENMOD displays the "Covariance Matrix (Model-Based)" and "Covariance Matrix (Empirical)" tables.

### GEE Parameter Estimate Correlation Matrices

If you specify the REPEATED statement and the CORRB option, PROC GENMOD displays the "Correlation Matrix (Model-Based)" and "Correlation Matrix (Empirical)" tables.

### GEE Working Correlation Matrix

If you specify the REPEATED statement and the CORRW option, PROC GENMOD displays the "Working Correlation Matrix" table.

### GEE Fit Criteria

If you specify the REPEATED statement, PROC GENMOD displays the quasi-likelihood information criteria for model fit QIC and $\mathrm{QIC}_u$ in the "GEE Fit Criteria" table.

## Analysis of GEE Parameter Estimates

If you specify the REPEATED statement, PROC GENMOD uses empirical standard error estimates to compute and display the "Analysis Of GEE Parameter Estimates Empirical Standard Error Estimates" table that contains the parameter names as follows:

- 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

In addition, the parameter estimate, the empirical standard error, a 95% confidence interval, and the $Z$ score and $p$-value are displayed for each parameter.

If you specify the MODELSE option in the REPEATED statement, the "Analysis Of GEE Parameter Estimates Model-Based Standard Error Estimates" 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
- the value of the linear predictor
- the standard error of the linear predictor
- confidence limits for the predicted values
- raw residual
- Pearson residual
- cluster number
- leverage
- cluster leverage
- cluster Cook's distance statistic
- studentized cluster Cook's distance statistic
- individual observation Cook's distance statistic
- cluster DFBETA statistic for each parameter
- cluster standardized DFBETA statistic for each parameter
- individual observation DFBETA statistic for each parameter
- individual observation standardized DFBETA statistic for each parameter

## Displayed Output for Bayesian Analysis

If a Bayesian analysis is requested with a BAYES statement, the displayed output includes the following.

### Model Information

The “Model Information” table displays the two-level data set name, the number of burn-in iterations, the number of iterations after the burn-in, the number of thinning iterations, the response distribution, the link function, the response variable name, the offset variable name, the frequency variable name, the scale weight variable name, the number of observations used, the number of events if events/trials format is used for response, the number of trials if events/trials format is used for response, the sum of frequency weights, the number of missing values in data set, and the 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

The “Class Level Information” table displays the levels of classification variables if you specify a CLASS statement.

### Maximum Likelihood Estimates

The “Analysis of Maximum Likelihood Parameter Estimates” table displays the maximum likelihood estimate of each parameter, the estimated standard error of the parameter estimator, and confidence limits for each parameter.

### Coefficient Prior

The “Coefficient Prior” table displays the prior distribution of the regression coefficients.

### Independent Prior Distributions for Model Parameters

The “Independent Prior Distributions for Model Parameters” table displays the prior distributions of additional model parameters (scale, exponential scale, Weibull scale, Weibull shape, gamma shape).

### Initial Values and Seeds

The “Initial Values and Seeds” table displays the initial values and random number generator seeds for the Gibbs chains.

### Fit Statistics

The “Fit Statistics” table displays the deviance information criterion (DIC) and the effective number of parameters.

### Descriptive Statistics of the Posterior Samples

The “Descriptive Statistics of the Posterior Sample” table contains the size of the sample, the mean, the standard deviation, and the quartiles for each model parameter.

### Interval Estimates for Posterior Sample

The “Interval Estimates for Posterior Sample” table contains the HPD intervals and the credible intervals for each model parameter.

### Correlation Matrix of the Posterior Samples

The “Correlation Matrix of the Posterior Samples” table is produced if you include the CORR suboption in the SUMMARY= option in the BAYES statement. This table displays the sample correlation of the posterior samples.

### Covariance Matrix of the Posterior Samples

The “Covariance Matrix of the Posterior Samples” table is produced if you include the COV suboption in the SUMMARY= option in the BAYES statement. This table displays the sample covariance of the posterior samples.

### Autocorrelations of the Posterior Samples

The “Autocorrelations of the Posterior Samples” table displays the lag1, lag5, lag10, and lag50 autocorrelations for each parameter.

### Gelman and Rubin Diagnostics

The “Gelman and Rubin Diagnostics” table is produced if you include the GELMAN suboption in the DIAGNOSTIC= option in the BAYES statement. This table displays the estimate of the potential scale reduction factor and its 97.5% upper confidence limit for each parameter.

### Geweke Diagnostics

The “Geweke Diagnostics” table displays the Geweke statistic and its *p*-value for each parameter.

### Raftery and Lewis Diagnostics

The “Raftery Diagnostics” tables is produced if you include the RAFTERY suboption in the DIAGNOSTIC= option in the BAYES statement. This table displays the Raftery and Lewis diagnostics for each variable.

### Heidelberger and Welch Diagnostics

The “Heidelberger and Welch Diagnostics” table is displayed if you include the HEIDELBERGER suboption in the DIAGNOSTIC= option in the BAYES statement. This table shows the results of a stationary test and a halfwidth test for each parameter.

### Effective Sample Size

The “Effective Sample Size” table displays, for each parameter, the effective sample size, the correlation time, and the efficiency.

### Monte Carlo Standard Errors

The “Monte Carlo Standard Errors” table displays, for each parameter, the Monte Carlo standard error, the posterior sample standard deviation, and the ratio of the two.

## Displayed Output for Exact Analysis

If an exact analysis is requested with an EXACT statement, the displayed output includes the following tables. If the METHOD=NETWORKMC option is specified, the test and estimate tables are renamed “Monte Carlo” tables and a Monte Carlo standard error column ($\sqrt{p(1-p)/n}$) is displayed.

### Sufficient Statistics

Displays if you request an OUTDIST= data set in an EXACT statement. The table lists the parameters and their observed sufficient statistics.

### (Monte Carlo) Conditional Exact Tests

This table tests the hypotheses that the parameters of interest are insignificant. See the section “Exact Logistic and Exact Poisson Regression” on page 3684 for details.

### (Monte Carlo) Exact Parameter Estimates

Displays if you specify the ESTIMATE option in the EXACT statement. This table gives individual parameter estimates for each variable (conditional on the values of all the other parameters in the model), confidence limits, and a two-sided $p$-value (twice the one-sided $p$-value) for testing that the parameter is zero. See the section “Exact Logistic and Exact Poisson Regression” on page 3684 for details.

### (Monte Carlo) Exact Odds Ratios

Displays if you specify the ESTIMATE=ODDS or ESTIMATE=BOTH option in the EXACT statement. See the section “Exact Logistic and Exact Poisson Regression” on page 3684 for details.

### Strata Summary

Displays if a STRATA statement is also specified. Shows the pattern of the number of events and the number of nonevents, or of the number of observations, in a stratum. See the section “STRATA Statement” on page 3637 for more information.

### Strata Information

Displays if a STRATA statement is specified with the INFO option.

## 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 separately in Table 51.13 for a maximum likelihood analysis, in Table 51.14 for a Bayesian analysis, and in Table 51.15 for an Exact analysis. For more information about ODS, see Chapter 23, “Using the Output Delivery System.”

**Table 51.13** ODS Tables Produced in PROC GENMOD for a Classical Analysis

| **ODS Table Name** | **Description** | **Statements** | **Option** |
|---|---|---|---|
| AssessmentSummary | Model assessment summary | ASSESS | Default |
| ClassLevels | Classification 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 |
| GEEExchCorr | GEE exchangeable working correlation value | REPEATED | TYPE=EXCH |
| GEEFitCriteria | GEE QIC fit criteria | 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 |

**Table 51.13** *continued*

| ODS Table Name | Description | Statements | Option |
|---|---|---|---|
| GEEWCorr | GEE working correlation matrix | REPEATED | CORRW |
| IterContrasts | Iteration history for contrasts | MODEL<br>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<br>REPEATED | ITPRINT |
| IterType3 | Iteration history for Type 3 statistics | MODEL | TYPE3 ITPRINT |
| LRCI | Likelihood ratio confidence intervals | MODEL | LRCI ITPRINT |
| Coef | Coefficients for least squares means | LSMEANS | E |
| Diffs | 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<br>REPEATED | ITPRINT |
| LastGradHess | Last evaluation of the gradient and Hessian | MODEL | ITPRINT |
| LinDep | Linearly dependent rows of contrasts | CONTRAST | Default |
| ModelANOVA | Type 3 tests or joint tests | MODEL | TYPE3 without REPEATED \| ZEROMODEL |
| ModelInfo | Model information | MODEL | Default |
| Modelfit | Goodness-of-fit statistics | MODEL | Default without REPEATED |
| NObs | Number of observations summary | | Default |
| NonEst | Nonestimable rows of contrasts | CONTRAST | Default |
| ObStats | Observation-wise statistics | MODEL | OBSTATS \| CL \| PREDICTED \| RESIDUALS \| XVARS |
| ParameterEstimates | Parameter estimates | MODEL | Default without REPEATED \| PRINTMLE with REPEATED |
| ParmInfo | Parameter indices | MODEL | Default |

**Table 51.13** *continued*

| ODS Table Name | Description | Statements | Option |
|---|---|---|---|
| ResponseProfile | Frequency counts for multinomial and binary models | MODEL | DIST=MULTINOMIAL \| DIST=BINOMIAL |
| Type1 | Type 1 tests | MODEL | TYPE1 |
| Type3 | Type 3 tests or joint tests for GEE model | MODEL<br>REPEATED | TYPE3 |
| Type3Zero | Type 3 tests or joint tests for zero-inflated model | MODEL<br>ZEROMODEL | TYPE3 |
| ZeroParameterEstimates | Parameter estimates for zero-inflated model | ZEROMODEL | Default |

**Table 51.14** ODS Tables Produced in PROC GENMOD for a Bayesian Analysis

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| AutoCorr | Autocorrelations of the posterior samples | BAYES | Default |
| ClassLevels | Classification variable levels | CLASS | Default |
| CoeffPrior | Prior distribution of the regression coefficients | BAYES | Default |
| ConvergenceStatus | Convergence status of maximum likelihood estimation | MODEL | Default |
| Corr | Correlation matrix of the posterior samples | BAYES | SUMMARY=CORR |
| ESS | Effective sample size | BAYES | Default |
| FitStatistics | Fit statistics | BAYES | Default |
| Gelman | Gelman and Rubin convergence diagnostics | BAYES | DIAG=GELMAN |
| Geweke | Geweke convergence diagnostics | BAYES | Default |
| Heidelberger | Heidelberger and Welch convergence diagnostics | BAYES | DIAG=HEIDELBERGER |
| InitialValues | Initial values of the Markov chains | BAYES | Default |
| IterParms | Iteration history for parameter estimates | MODEL | ITPRINT |
| LastGradHess | Last evaluation of the gradient and Hessian for maximum likelihood estimation | MODEL | ITPRINT |
| MCSE | Monte Carlo standard errors | BAYES | DIAG=MCSE |
| ModelInfo | Model information | PROC | Default |

**Table 51.14** *continued*

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| NObs | Number of observations | | Default |
| ParameterEstimates | Maximum likelihood estimates of model parameters | MODEL | Default |
| ParmInfo | Parameter indices | MODEL | Default |
| ParmPrior | Prior distribution for scale and shape | BAYES | Default |
| PostIntervals | HPD and equal-tail intervals of the posterior samples | BAYES | Default |
| PosteriorSample | Posterior samples (for ODS output data set only) | BAYES | |
| PostSummaries | Summary statistics of the posterior samples | BAYES | Default |
| Raftery | Raftery and Lewis convergence diagnostics | BAYES | DIAG=RAFTERY |

**Table 51.15** ODS Tables Produced in PROC GENMOD for an Exact Analysis

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| ExactParmEst | Parameter estimates | EXACT | ESTIMATE,<br>ESTIMATE=PARM,<br>ESTIMATE=BOTH |
| ExactTests | Conditional exact tests | EXACT | Default |
| ExpExactParmEst | Exact odds ratios | EXACT | ESTIMATE=ODDS,<br>ESTIMATE=BOTH |
| NStrataIgnored | Number of uninformative strata | STRATA | Default |
| StrataSummary | Number of strata with specific response frequencies | STRATA | Default |
| StrataInfo | Event and nonevent frequencies for each stratum | STRATA | INFO |
| SuffStats | Sufficient statistics | EXACT | OUTDIST= |

## ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 24, "Statistical Graphics Using ODS."

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 687 in Chapter 24, "Statistical Graphics Using ODS."

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section "A Primer on ODS Statistical Graphics" on page 686 in Chapter 24, "Statistical Graphics Using ODS."

Some graphs are produced by default; other graphs are produced by using statements and options. You can reference every graph produced through ODS Graphics with a name. The names of the graphs that PROC GENMOD generates are listed in Table 51.16, along with the required statements and options.

### ODS Graph Names

PROC GENMOD assigns a name to each graph it creates using ODS. You can use these names to reference the graphs when using ODS. The names are listed in Table 51.16.

To request these graphs, ODS Graphics must be enabled and you must specify the statement and options indicated in Table 51.16.

**Table 51.16** Graphs Produced by PROC GENMOD

| ODS Graph Name | Description | Statement | Option |
|---|---|---|---|
| ADPanel | Autocorrelation function and density panel | BAYES | PLOTS=(AUTOCORR DENSITY) |
| AutocorrPanel | Autocorrelation function panel | BAYES | PLOTS= AUTOCORR |
| AutocorrPlot | Autocorrelation function plot | BAYES | PLOTS(UNPACK)=AUTOCORR |
| ClusterCooksDPlot | Cluster Cook's *D* by cluster number | PROC | PLOTS= |
| ClusterDFFITPlot | Cluster DFFIT by cluster number | PROC | PLOTS= |
| ClusterLeveragePlot | Cluster leverage by cluster number | PROC | PLOTS= |
| CooksDPlot | Cook's distance | PROC | PLOTS= |
| CumResidPanel | Panel of aggregates of residuals | ASSESS | CRPANEL |
| CumulativeResiduals | Model assessment based on aggregates of residuals | ASSESS | Default |
| DevianceResidByXBeta | Deviance residuals by linear predictor | PROC | PLOTS= |
| DevianceResidualPlot | Deviance values | PROC | PLOTS= |

**Table 51.16** *continued*

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| DFBETAByCluster | Cluster DFBeta by cluster number | PROC | PLOTS= |
| DFBETAPlot | DFBeta | PROC | PLOTS= |
| DiagnosticPlot | Panel of residuals, influence, and diagnostic statistics | PROC<br>MODEL<br>REPEATED | PLOTS= |
| LeveragePlot | Leverage | PROC | PLOTS= |
| LikeResidByXBeta | Likelihood residuals by linear predictor | PROC | PLOTS= |
| LikeResidualPlot | Likelihood residuals | PROC | PLOTS= |
| PearsonResidByXBeta | Pearson residuals by linear predictor | PROC | PLOTS= |
| PearsonResidualPlot | Pearson residuals | PROC | PLOTS= |
| PredictedByObservation | Predicted values | PROC | PLOTS= |
| RawResidByXBeta | Raw residuals by linear predictor | PROC | PLOTS= |
| RawResidualPlot | Raw residuals | PROC | PLOTS= |
| StdDevianceResidByXBeta | Standardized deviance residuals by linear predictor | PROC | PLOTS= |
| StdDevianceResidualPlot | Standardized deviance residuals | PROC | PLOTS= |
| StdDFBETAByCluster | Standardized cluster DFBeta by cluster number | PROC | PLOTS= |
| StdDFBETAPlot | Standardized DFBeta | PROC | PLOTS= |
| StdPearsonResidByXBeta | Standardized Pearson residuals by linear predictor | PROC | PLOTS= |
| StdPearsonResidualPlot | Standardized Pearson residuals | PROC | PLOTS= |
| TAPanel | Trace and autocorrelation function panel | BAYES | PLOTS=(TRACE AUTOCORR) |
| TADPanel | Trace, autocorrelation, and density function panel | BAYES | Default |
| TDPanel | Trace and density panel | BAYES | PLOTS=(TRACE DENSITY) |
| TracePanel | Trace panel | BAYES | PLOTS=TRACE |
| TracePlot | Trace plot | BAYES | PLOTS(UNPACK)=TRACE |
| ZeroInflationProbPlot | Zero-inflation probabilities | PROC | PLOTS= |

# Examples: GENMOD Procedure

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 for guidance on complete analysis of data by using generalized linear models.

## Example 51.1: 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 51.1.1.

**Output 51.1.1** Model Information

**The GENMOD Procedure**

| Model Information | |
|---|---|
| **Data Set** | WORK.DRUG |
| **Distribution** | Binomial |
| **Link Function** | Logit |
| **Response Variable (Events)** | r |
| **Response Variable (Trials)** | n |

The five levels of the CLASS variable DRUG are displayed in Output 51.1.2.

**Output 51.1.2** CLASS Variable Levels

| 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 51.1.3, 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 51.1.3** Goodness-of-Fit Criteria

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 | |
| **Full Log Likelihood** | | -23.7343 | |
| **AIC (smaller is better)** | | 59.4686 | |
| **AICC (smaller is better)** | | 67.1050 | |
| **BIC (smaller is better)** | | 64.8109 | |

In the "Analysis Of Parameter Estimates" table displayed in Output 51.1.4, 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 section "Overdispersion" on page 3650 for a discussion of overdispersion.

**Output 51.1.4** Parameter Estimates

**Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | | DF | Estimate | Standard Error | Likelihood Ratio 95% Confidence Limits | | Wald 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 (see 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 3655 for a discussion of profile likelihood confidence intervals.

## Example 51.2: 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;
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 results from these statements are displayed in Output 51.2.1.

**Output 51.2.1** Log-Linked Normal Regression

**The GENMOD Procedure**

| Model Information | |
|---|---|
| **Data Set** | WORK.NOR |
| **Distribution** | Normal |
| **Link Function** | Log |
| **Dependent Variable** | y |

**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 | |
| **Full Log Likelihood** | | -32.1783 | |
| **AIC (smaller is better)** | | 70.3566 | |
| **AICC (smaller is better)** | | 72.3566 | |
| **BIC (smaller is better)** | | 72.6743 | |

**Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Wald 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.

```
proc print data=Residuals;
run;
```

**Output 51.2.2** Data Set of Predicted Values and Residuals

| Obs | x | y | Pred | Reschi | Resraw | Resdev | 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 51.2.2) is created by the OUTPUT statement. You can use the PLOTS= option in the PROC GENMOD statement to create plots of predicted values and residuals. 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 51.3: 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. See Lawless (2003) 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 51.3.1.

**Output 51.3.1** Gamma Model of Life Data

```
                    The GENMOD Procedure

                     Model Information
            Data Set                 WORK.LIFDAT
            Distribution                   Gamma
            Link Function                    Log
            Dependent Variable          lifetime

                       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
    Full Log Likelihood             -1432.4177
    AIC (smaller is better)          2870.8353
    AICC (smaller is better)         2870.9572
    BIC (smaller is better)          2880.7453

          Analysis Of Maximum Likelihood Parameter Estimates
                                         Wald 95%
                             Standard   Confidence        Wald
Parameter     DF  Estimate      Error     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 between 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 51.3.2 displays the results of fitting the model with the mfg effect omitted.

**Output 51.3.2** Refitting of the Gamma Model: Omitting the mfg Effect

**The GENMOD Procedure**

**Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | Likelihood Ratio 95% Confidence Limits | | Wald Chi-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 section "Response Probability Distributions" on page 3640 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 51.4: 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 of the three brands. The variable taste contains the ratings, and the variable 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 data=Icecream 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 $\gamma_j(\mathrm{x}) = \Pr(\mathrm{taste} \le j)$ is the cumulative probability of the *j*th or lower taste category, then the odds ratio comparing $\mathrm{x}_1$ to $\mathrm{x}_2$ is as follows:

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

See 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 51.4.1 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 51.4.1** 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 |

**Class Level Information**

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

**Response Profile**

| Ordered Value | taste | Total Frequency |
|---|---|---|
| **1** | vg | 140 |
| **2** | g | 162 |
| **3** | m | 421 |
| **4** | b | 156 |
| **5** | vb | 166 |

Output 51.4.2 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 51.4.1. 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 51.4.2** Parameter Estimates

**Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|
| **Intercept1** | | 1 | -1.8578 | 0.1219 | -2.0967 | -1.6189 | 232.35 | <.0001 |
| **Intercept2** | | 1 | -0.8646 | 0.1056 | -1.0716 | -0.6576 | 67.02 | <.0001 |
| **Intercept3** | | 1 | 0.9231 | 0.1060 | 0.7154 | 1.1308 | 75.87 | <.0001 |
| **Intercept4** | | 1 | 1.8078 | 0.1191 | 1.5743 | 2.0413 | 230.32 | <.0001 |
| **brand** | ice1 | 1 | 0.3847 | 0.1370 | 0.1162 | 0.6532 | 7.89 | 0.0050 |
| **brand** | ice2 | 1 | -0.6457 | 0.1397 | -0.9196 | -0.3719 | 21.36 | <.0001 |
| **brand** | ice3 | 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 Type 1 test displayed in Output 51.4.3 indicates that Brand is highly significant; that is, there are significant differences among the brands. The log odds ratios and odds ratios in the "ESTIMATE Statement

Results” table indicate the relative differences among 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 51.4.3** Type 1 Tests and Odds Ratios

**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 | Mean Estimate | Mean Confidence Limits | | L'Beta Estimate | Standard Error | Alpha | L'Beta Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|---|---|
| **LogOR12** | 0.7370 | 0.6805 | 0.7867 | 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** | 0.5950 | 0.5290 | 0.6577 | 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** | 0.3439 | 0.2850 | 0.4081 | -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 51.5: GEE for Binary Data with Logit Link Function

Output 51.5.1 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 (2000).

Patients in each of two centers are randomly assigned to groups receiving the active treatment or a placebo. During treatment, respiratory status, represented by the variable outcome (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 ($x_{ij5}$), and baseline ($x_{ij6}$), so that $x' = [x_{ij1}, x_{ij2}, \ldots, x_{ij6}]$ 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. To be consistent with the analysis in Stokes, Davis, and Koch (2000), the four classification explanatory variables are coded as follows via options in the CLASS statement:

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

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

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.

**Output 51.5.1** Respiratory Disorder Data

| Obs | center | id | treatment | sex | age | baseline | visit1 | visit2 | visit3 | visit4 | visit | outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | P | M | 46 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 2 | 1 | 1 | P | M | 46 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 3 | 1 | 1 | P | M | 46 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| 4 | 1 | 1 | P | M | 46 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 5 | 1 | 2 | P | M | 28 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 6 | 1 | 2 | P | M | 28 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 7 | 1 | 2 | P | M | 28 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| 8 | 1 | 2 | P | M | 28 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| . | | | | | | | | | | | | |
| . | | | | | | | | | | | | |
| . | | | | | | | | | | | | |
| 214 | 2 | 1 | P | F | 39 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 215 | 2 | 1 | P | F | 39 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 216 | 2 | 1 | P | F | 39 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| 217 | 2 | 1 | P | F | 39 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 218 | 2 | 2 | A | M | 25 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 219 | 2 | 2 | A | M | 25 | 0 | 0 | 1 | 1 | 1 | 2 | 1 |
| 220 | 2 | 2 | A | M | 25 | 0 | 0 | 1 | 1 | 1 | 3 | 1 |
| 221 | 2 | 2 | A | M | 25 | 0 | 0 | 1 | 1 | 1 | 4 | 1 |
| . | | | | | | | | | | | | |
| . | | | | | | | | | | | | |
| . | | | | | | | | | | | | |
| . | | | | | | | | | | | | |

The GEE solution is requested with the REPEATED statement in the GENMOD procedure. The option SUBJECT=ID(CENTER) specifies that the observations in any single cluster are uniquely identified by both center and id. An equivalent specification is SUBJECT=ID*CENTER. Since the same id values are used in each center, one of these specifications is needed. If id values were unique across all centers, SUBJECT=ID would be specified.

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. The following SAS statements perform the GEE model fit:

```
proc genmod data=resp;
   class id treatment(ref="P") center(ref="1") sex(ref="M")
      baseline(ref="0");
   model outcome(event='1')=treatment center sex age baseline / dist=bin;
   repeated subject=id(center) / corr=unstr corrw;
run;
```

These statements first fit the generalized linear (GLM) model specified in the MODEL statement. The parameter estimates from the generalized linear model fit are not shown in the output, but they are used as initial values for the GEE solution. The EVENT='1' option in the MODEL statement models the probability that outcome = 1. If the EVENT='1' option had not been specified, the probability that outcome = 0 would be modeled by default.

Information about the GEE model is displayed in Output 51.5.2. The results of GEE model fitting are displayed in Output 51.5.3. Model goodness-of-fit criteria are displayed in Output 51.5.4. 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 51.5.2** 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 51.5.3** Results of Model Fitting

**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 |
| **treatment** | A | 1.2442 | 0.3455 | 0.5669 | 1.9214 | 3.60 | 0.0003 |
| **treatment** | P | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **center** | 2 | 0.6558 | 0.3512 | -0.0326 | 1.3442 | 1.87 | 0.0619 |
| **center** | 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **sex** | F | 0.1128 | 0.4408 | -0.7512 | 0.9768 | 0.26 | 0.7981 |
| **sex** | M | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **age** | | -0.0175 | 0.0129 | -0.0427 | 0.0077 | -1.36 | 0.1728 |
| **baseline** | 1 | 1.8981 | 0.3441 | 1.2237 | 2.5725 | 5.52 | <.0001 |
| **baseline** | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |

**Output 51.5.4** Model Fit Criteria

| GEE Fit Criteria | |
|---|---|
| QIC | 512.3416 |
| QICu | 499.6081 |

The nonsignificance of age and sex make them candidates for omission from the model.

## Example 51.6: Log Odds Ratios and the ALR Algorithm

Since the respiratory data in Example 51.5 are binary, you can use the ALR algorithm to model the log odds ratios instead of using working correlations to model associations. In this example, 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 51.5 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 treatment(ref="P") center(ref="1") sex(ref="M")
      baseline(ref="0");
   model outcome(event='1')=treatment center sex age baseline / dist=bin;
   repeated  subject=id(center) / logor=fullclust;
run;
```

The results of fitting the model are displayed in Output 51.6.1 along with a table that shows the correspondence between the log odds ratio parameters and the within-cluster pairs. Model goodness-of-fit criteria are shown in Output 51.6.2. The QIC for the ALR model shown in Output 51.6.2 is 511.86, whereas the QIC for the unstructured working correlation model shown in Output 51.5.4 is 512.34, indicating that the ALR model is a slightly better fit.

**Output 51.6.1** 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) |

**Output 51.6.1** *continued*

**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 |
| **treatment** | A | 1.2611 | 0.3406 | 0.5934 | 1.9287 | 3.70 | 0.0002 |
| **treatment** | P | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **center** | 2 | 0.6287 | 0.3486 | -0.0545 | 1.3119 | 1.80 | 0.0713 |
| **center** | 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **sex** | F | 0.1024 | 0.4362 | -0.7526 | 0.9575 | 0.23 | 0.8144 |
| **sex** | M | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **age** | | -0.0162 | 0.0125 | -0.0407 | 0.0084 | -1.29 | 0.1977 |
| **baseline** | 1 | 1.8980 | 0.3404 | 1.2308 | 2.5652 | 5.58 | <.0001 |
| **baseline** | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **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 |

**Output 51.6.2** Model Fit Criteria

| GEE Fit Criteria | |
|---|---|
| QIC | 511.8589 |
| QICu | 499.6516 |

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

**Output 51.6.3** 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 treatment(ref="P") center(ref="1") sex(ref="M")
         baseline(ref="0");
      model outcome(event='1')=treatment center sex age baseline / dist=bin;
      repeated  subject=id(center) / logor=zfull
                                     zdata=zin
                                     zrow =(z1-z6)
                                     ypair=(y1 y2);
   run;
```

## Example 51.7: Log-Linear Model for Count Data

In this example the data, from Thall and Vail (1990), concern 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 51.17. See "Gee Model for Count Data, Exchangeable Correlation" in the SAS/STAT Sample Program Library for the complete data set.

**Table 51.17** 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

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

where

$$Y_{ij} = \text{number of epileptic seizures in interval } j$$

$$t_{ij} = \text{length of interval } j$$

$$x_{i1} = \begin{cases} 1: & \text{weeks 8–16 (treatment)} \\ 0: & \text{weeks 0–8 (baseline)} \end{cases}$$

$$x_{i2} = \begin{cases} 1: & \text{progabide group} \\ 0: & \text{placebo group} \end{cases}$$

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 51.18.

**Table 51.18** 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.

Output 51.7.1 lists the first 14 observations of the data, which are arranged as one visit per observation:

**Output 51.7.1** Partial Listing of the Seizure Data

| Obs | id | y | visit | trt | bline | age |
|---|---|---|---|---|---|---|
| **1** | 104 | 5 | 1 | 0 | 11 | 31 |
| **2** | 104 | 3 | 2 | 0 | 11 | 31 |
| **3** | 104 | 3 | 3 | 0 | 11 | 31 |
| **4** | 104 | 3 | 4 | 0 | 11 | 31 |
| **5** | 106 | 3 | 1 | 0 | 11 | 30 |
| **6** | 106 | 5 | 2 | 0 | 11 | 30 |
| **7** | 106 | 3 | 3 | 0 | 11 | 30 |
| **8** | 106 | 3 | 4 | 0 | 11 | 30 |
| **9** | 107 | 2 | 1 | 0 | 6 | 25 |
| **10** | 107 | 4 | 2 | 0 | 6 | 25 |
| **11** | 107 | 0 | 3 | 0 | 6 | 25 |
| **12** | 107 | 5 | 4 | 0 | 6 | 25 |
| **13** | 114 | 4 | 1 | 0 | 8 | 36 |
| **14** | 114 | 4 | 2 | 0 | 8 | 36 |

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. The following statements prepare the data for analysis with PROC GENMOD:

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

data new;
   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;
```

For comparison with the GEE results, an ordinary Poisson regression is first fit. The results are shown in Output 51.7.2.

**Output 51.7.2** Maximum Likelihood Estimates

**The GENMOD Procedure**

**Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Wald 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.

The GEE solution is requested with the REPEATED statement in the GENMOD procedure. The SUBJECT=ID option indicates that the variable id describes the observations 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.

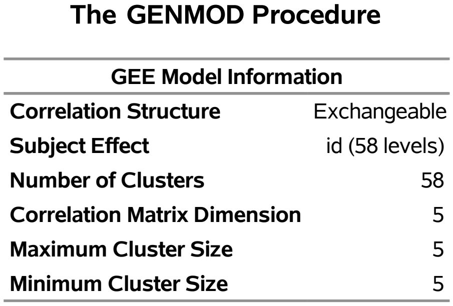
The following statements perform the analysis:

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

These statements first fit a generalized linear model (GLM) to these data by maximum likelihood. The estimates are not shown in the output, but are used as initial values for the GEE solution.

Information about the GEE model is displayed in Output 51.7.3. The results of fitting the model are displayed in Output 51.7.4. Compare these with the model of independence displayed in Output 51.7.2. 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.

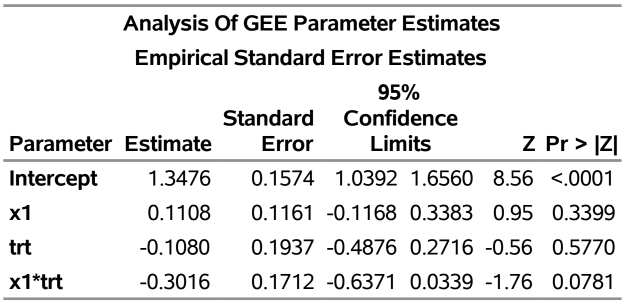
**Output 51.7.3** GEE Model Information



**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 51.7.4** GEE Parameter Estimates



**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 51.19 displays the regression coefficients, standard errors, and normalized coefficients that result from fitting the model with independent and exchangeable working correlation matrices.

**Table 51.19** 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 51.7.5.

**Output 51.7.5** Working Correlation Matrix

**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 51.7.6 contains these estimates.

**Output 51.7.6** Covariance Matrices

**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.

## Example 51.8: Model Assessment of Multiple Regression Using Aggregates of Residuals

This example illustrates the use of cumulative residuals to assess the adequacy of a normal linear regression model.

Neter et al. (1996, Section 8.2) describe a study of 54 patients undergoing a certain kind of liver operation in a surgical unit. The data consist of the survival time and certain covariates. After a model selection procedure, they arrived at the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where $Y$ is the logarithm (base 10) of the survival time; $X_1$, $X_2$, $X_3$ are *blood-clotting score*, *prognostic index*, and *enzyme function*, respectively; and $\epsilon$ is a normal error term. A listing of the SAS data set containing the data is shown in Output 51.8.1. The variables Y, X1, X2, and X3 correspond to $Y$, $X_1$, $X_2$, and $X_3$, and LogX1 is $\log(X_1)$. The PROC GENMOD fit of the model is shown in Output 51.8.2. The analysis first focuses on the adequacy of the functional form of $X_1$, *blood-clotting score*.

**Output 51.8.1** Surgical Unit Example Data

| Obs | Y | X1 | X2 | X3 | LogX1 |
|---:|---:|---:|---:|---:|---:|
| **1** | 2.3010 | 6.7 | 62 | 81 | 0.82607 |
| **2** | 2.0043 | 5.1 | 59 | 66 | 0.70757 |
| **3** | 2.3096 | 7.4 | 57 | 83 | 0.86923 |
| **4** | 2.0043 | 6.5 | 73 | 41 | 0.81291 |
| **5** | 2.7067 | 7.8 | 65 | 115 | 0.89209 |
| **6** | 1.9031 | 5.8 | 38 | 72 | 0.76343 |
| **7** | 1.9031 | 5.7 | 46 | 63 | 0.75587 |
| **8** | 2.1038 | 3.7 | 68 | 81 | 0.56820 |
| **9** | 2.3054 | 6.0 | 67 | 93 | 0.77815 |
| **10** | 2.3075 | 3.7 | 76 | 94 | 0.56820 |
| **11** | 2.5172 | 6.3 | 84 | 83 | 0.79934 |
| **12** | 1.8129 | 6.7 | 51 | 43 | 0.82607 |
| **13** | 2.9191 | 5.8 | 96 | 114 | 0.76343 |
| **14** | 2.5185 | 5.8 | 83 | 88 | 0.76343 |
| **15** | 2.2253 | 7.7 | 62 | 67 | 0.88649 |
| **16** | 2.3365 | 7.4 | 74 | 68 | 0.86923 |
| **17** | 1.9395 | 6.0 | 85 | 28 | 0.77815 |
| **18** | 1.5315 | 3.7 | 51 | 41 | 0.56820 |
| **19** | 2.3324 | 7.3 | 68 | 74 | 0.86332 |
| **20** | 2.2355 | 5.6 | 57 | 87 | 0.74819 |
| **21** | 2.0374 | 5.2 | 52 | 76 | 0.71600 |
| **22** | 2.1335 | 3.4 | 83 | 53 | 0.53148 |
| **23** | 1.8451 | 6.7 | 26 | 68 | 0.82607 |
| **24** | 2.3424 | 5.8 | 67 | 86 | 0.76343 |
| **25** | 2.4409 | 6.3 | 59 | 100 | 0.79934 |
| **26** | 2.1584 | 5.8 | 61 | 73 | 0.76343 |
| **27** | 2.2577 | 5.2 | 52 | 86 | 0.71600 |
| **28** | 2.7589 | 11.2 | 76 | 90 | 1.04922 |
| **29** | 1.8573 | 5.2 | 54 | 56 | 0.71600 |
| **30** | 2.2504 | 5.8 | 76 | 59 | 0.76343 |
| **31** | 1.8513 | 3.2 | 64 | 65 | 0.50515 |
| **32** | 1.7634 | 8.7 | 45 | 23 | 0.93952 |
| **33** | 2.0645 | 5.0 | 59 | 73 | 0.69897 |
| **34** | 2.4698 | 5.8 | 72 | 93 | 0.76343 |
| **35** | 2.0607 | 5.4 | 58 | 70 | 0.73239 |
| **36** | 2.2648 | 5.3 | 51 | 99 | 0.72428 |
| **37** | 2.0719 | 2.6 | 74 | 86 | 0.41497 |
| **38** | 2.0792 | 4.3 | 8 | 119 | 0.63347 |
| **39** | 2.1790 | 4.8 | 61 | 76 | 0.68124 |
| **40** | 2.1703 | 5.4 | 52 | 88 | 0.73239 |
| **41** | 1.9777 | 5.2 | 49 | 72 | 0.71600 |
| **42** | 1.8751 | 3.6 | 28 | 99 | 0.55630 |
| **43** | 2.6840 | 8.8 | 86 | 88 | 0.94448 |
| **44** | 2.1847 | 6.5 | 56 | 77 | 0.81291 |
| **45** | 2.2810 | 3.4 | 77 | 93 | 0.53148 |
| **46** | 2.0899 | 6.5 | 40 | 84 | 0.81291 |
| **47** | 2.4928 | 4.5 | 73 | 106 | 0.65321 |
| **48** | 2.5999 | 4.8 | 86 | 101 | 0.68124 |

**Output 51.8.1** *continued*

| Obs | Y | X1 | X2 | X3 | LogX1 |
|---|---|---|---|---|---|
| **49** | 2.1987 | 5.1 | 67 | 77 | 0.70757 |
| **50** | 2.4914 | 3.9 | 82 | 103 | 0.59106 |
| **51** | 2.0934 | 6.6 | 77 | 46 | 0.81954 |
| **52** | 2.0969 | 6.4 | 85 | 40 | 0.80618 |
| **53** | 2.2967 | 6.4 | 59 | 85 | 0.80618 |
| **54** | 2.4955 | 8.8 | 78 | 72 | 0.94448 |

In order to assess the adequacy of the fitted multiple regression model, the ASSESS statement in the following SAS statements is used to create the plots of cumulative residuals against X1 shown in Output 51.8.3 and Output 51.8.4 and the summary table in Output 51.8.5:

```
ods graphics on;

proc genmod data=Surg;
   model Y = X1 X2 X3 / scale=Pearson;
   assess var=(X1) / resample=10000
                     seed=603708000
                     crpanel;
run;
```

**Output 51.8.2** Regression Model for Linear X1

**The GENMOD Procedure**

**Analysis Of Maximum Likelihood Parameter Estimates**

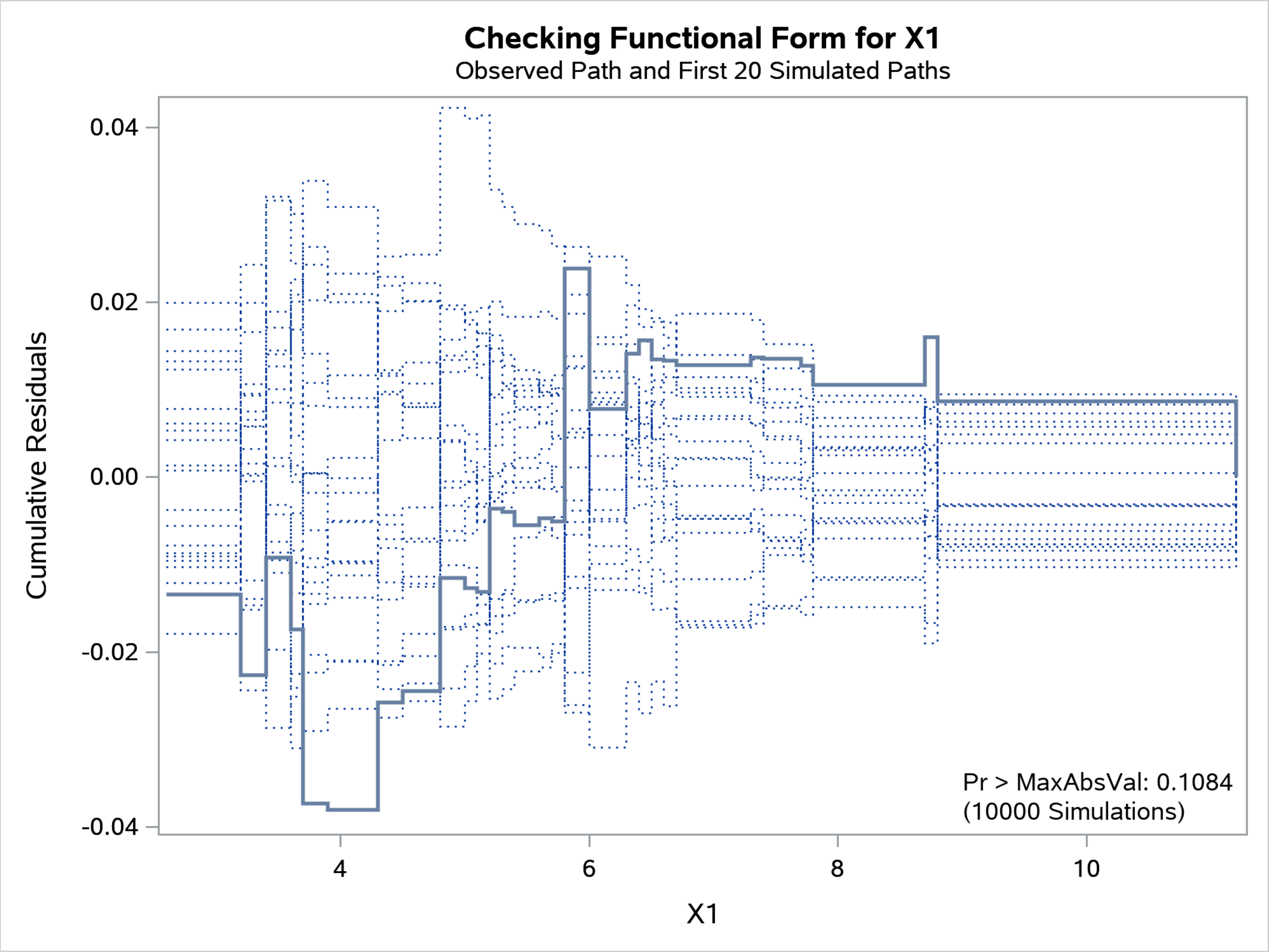
| | | | | Wald 95% | | | |
|---|---|---|---|---|---|---|---|
| | | | Standard | Confidence | | Wald | |
| **Parameter** | **DF** | **Estimate** | **Error** | **Limits** | | **Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | 0.4836 | 0.0426 | 0.4001 | 0.5672 | 128.71 | <.0001 |
| **X1** | 1 | 0.0692 | 0.0041 | 0.0612 | 0.0772 | 288.17 | <.0001 |
| **X2** | 1 | 0.0093 | 0.0004 | 0.0085 | 0.0100 | 590.45 | <.0001 |
| **X3** | 1 | 0.0095 | 0.0003 | 0.0089 | 0.0101 | 966.07 | <.0001 |
| **Scale** | 0 | 0.0469 | 0.0000 | 0.0469 | 0.0469 | | |

**Note:** The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

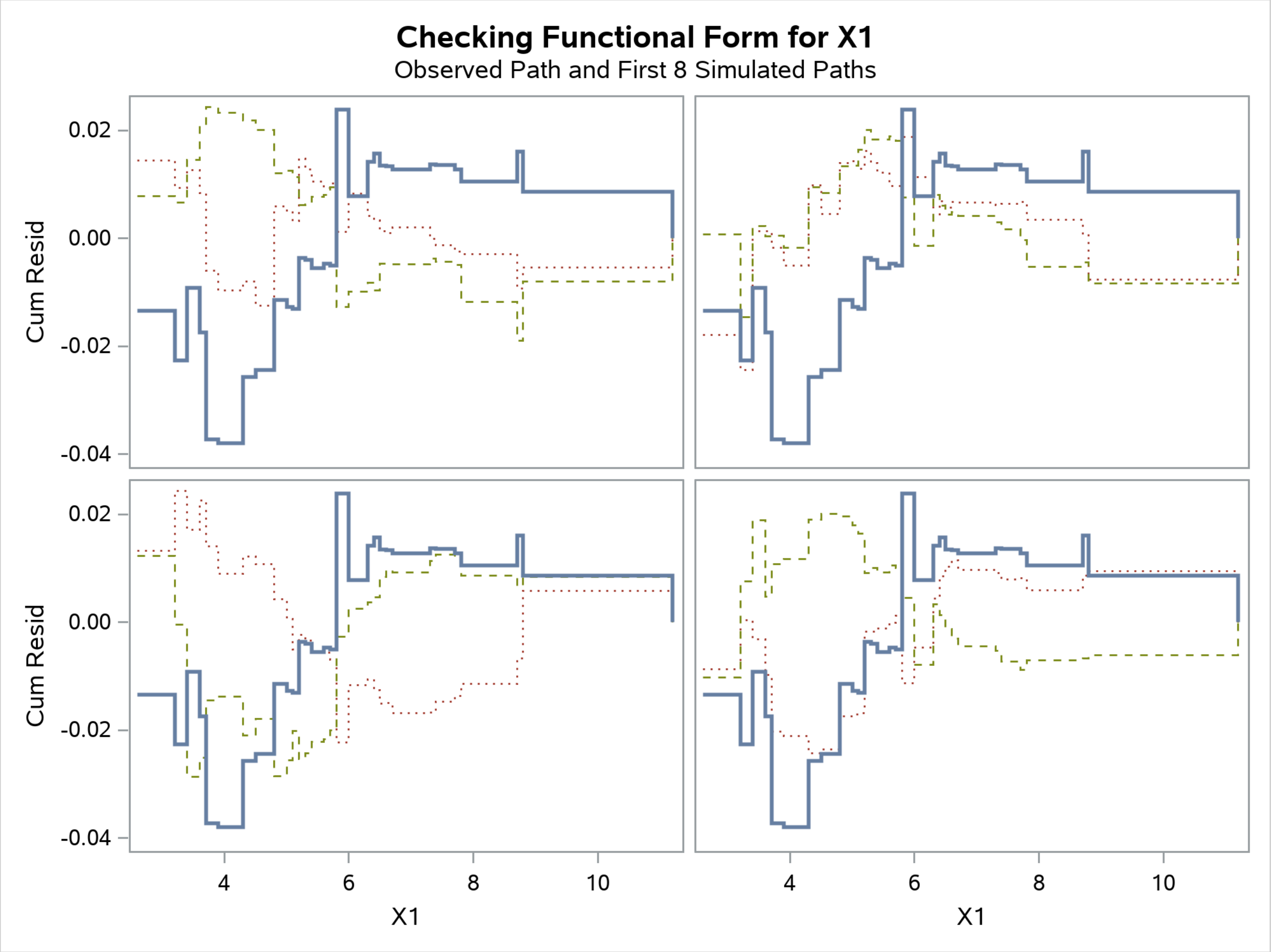
See Lin, Wei, and Ying (2002) for details about model assessment that uses cumulative residual plots. The RESAMPLE= keyword specifies that a *p*-value be computed based on a sample of 10,000 simulated residual paths. A random number seed is specified by the SEED= keyword for reproducibility. If you do not specify the seed, one is derived from the time of day. The keyword CRPANEL specifies that the panel of four cumulative residual plots shown in Output 51.8.4 be created, each with two simulated paths. The single residual plot with 20 simulated paths in Output 51.8.3 is created by default.

To request these graphs, ODS Graphics must be enabled and you must specify the ASSESS statement. For general information about ODS Graphics, see Chapter 24, "Statistical Graphics Using ODS." For specific information about the graphics available in the GENMOD procedure, see the section "ODS Graphics" on page 3703.

**Output 51.8.3** Cumulative Residual Plot for Linear X1 Fit



**Output 51.8.4** Cumulative Residual Panel Plot for Linear X1 Fit



**Output 51.8.5** Summary of Model Assessment

| Assessment Summary | | | | |
|---|---|---|---|---|
| Assessment Variable | Maximum Absolute Value | Replications | Seed | Pr > MaxAbsVal |
| X1 | 0.0380 | 10000 | 603708000 | 0.1084 |

The *p*-value of 0.1084 reported on Output 51.8.3 and Output 51.8.5 suggests that a more adequate model might be possible. The observed cumulative residuals in Output 51.8.3 and Output 51.8.4, represented by the heavy lines, seem atypical of the simulated curves, represented by the light lines, reinforcing the conclusion that a more appropriate functional form for X1 is possible.

The cumulative residual plots in Output 51.8.6 provide guidance in determining a more appropriate functional form. The four curves were created from simple forms of model misspecification by using simulated data. The mean models of the data and the fitted model are shown in Table 51.20.

**Output 51.8.6** Typical Cumulative Residual Patterns

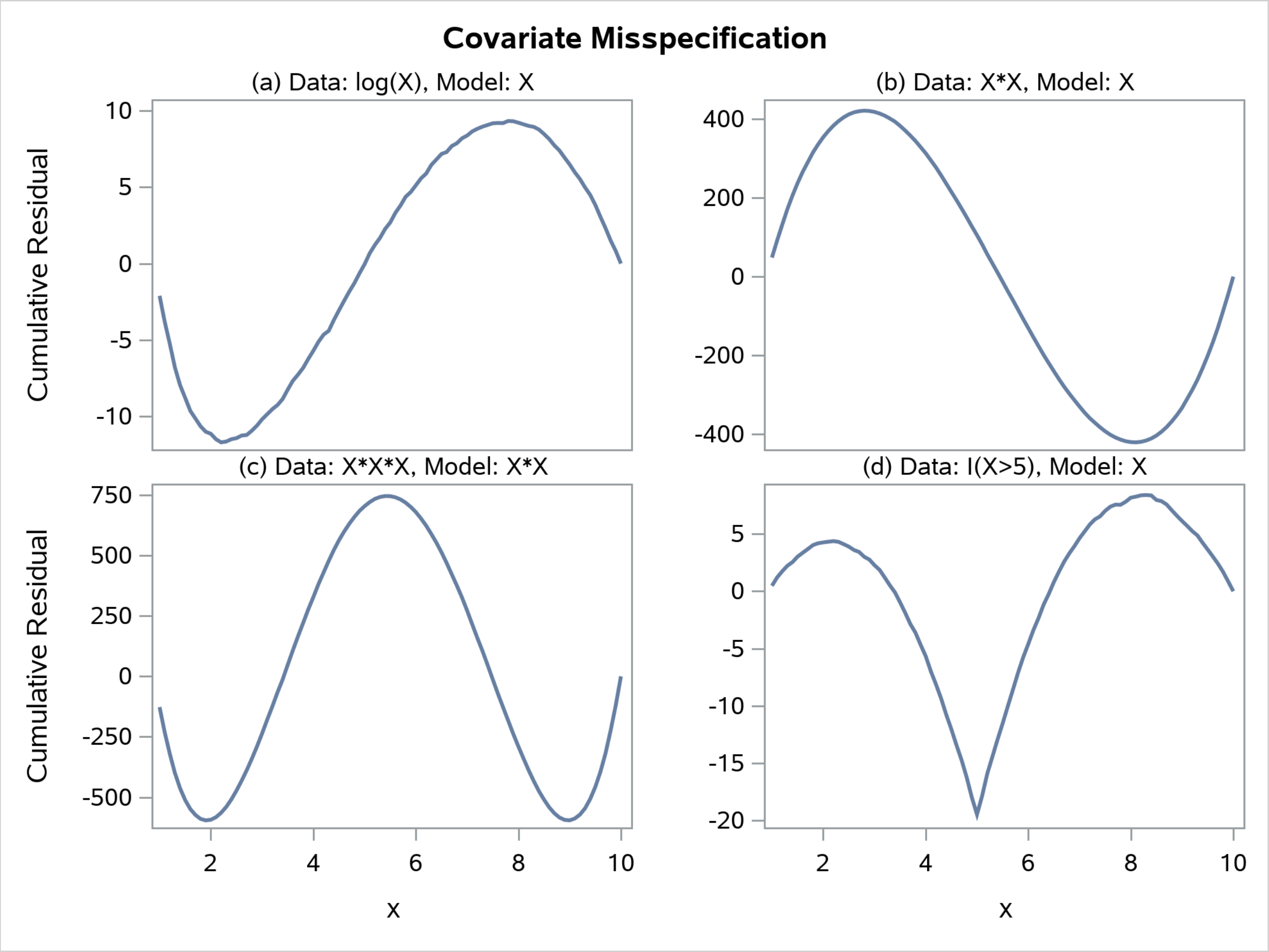


**Table 51.20** Model Misspecifications

| Plot | Data E($Y$) | Fitted Model E($Y$) |
|---|---|---|
| (a) | $\log(X)$ | $X$ |
| (b) | $X + X^2$ | $X$ |
| (c) | $X + X^2 + X^3$ | $X + X^2$ |
| (d) | $I(X > 5)$ | $X$ |

The observed cumulative residual pattern in Output 51.8.3 and Output 51.8.4 most resembles the behavior of the curve in plot (a) of Output 51.8.6, indicating that $\log(X_1)$ might be a more appropriate term in the model than $X_1$.

The following SAS statements fit a model with LogX1 in place of X1 and request a model assessment:

```
proc genmod data=Surg;
   model Y = LogX1 X2 X3 / scale=Pearson;
   assess var=(LogX1) / resample=10000
                        seed=603708000;
run;
```

The revised model fit is shown in Output 51.8.7, the *p*-value from the simulation is 0.4777, and the cumulative residuals plotted in Output 51.8.8 show no systematic trend. The log transformation for X1 is more appropriate. Under the revised model, the *p*-values for testing the functional forms of X2 and X3 are 0.20 and 0.63, respectively; and the *p*-value for testing the linearity of the model is 0.65. Thus, the revised model seems reasonable.

**Output 51.8.7** Multiple Regression Model with Log(X1)

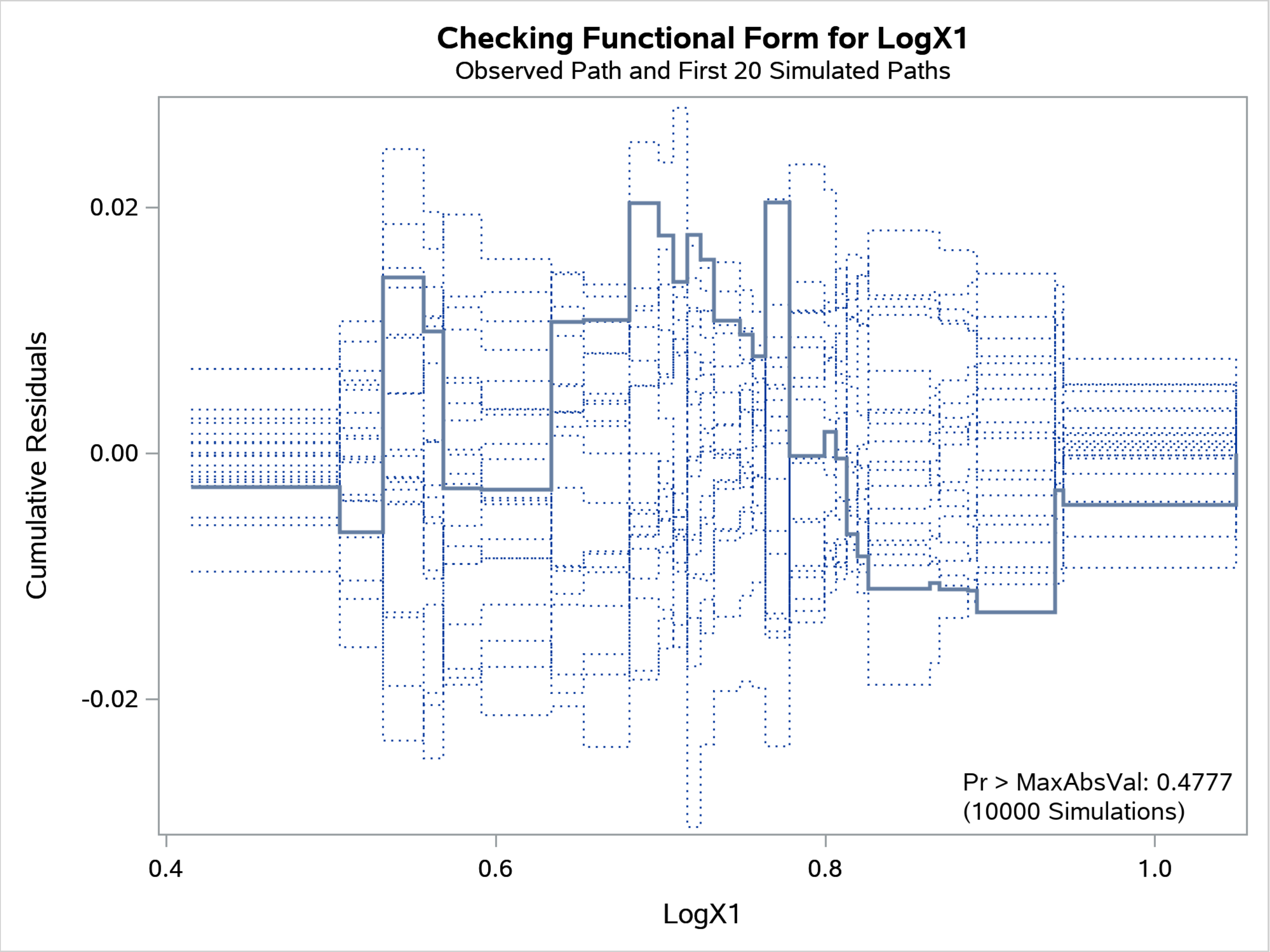
**The GENMOD Procedure**

**Analysis Of Maximum Likelihood Parameter Estimates**

| | | | | Wald 95% | | | |
|---|---|---|---|---|---|---|---|
| | | | Standard | Confidence | | Wald | |
| Parameter | DF | Estimate | Error | Limits | | Chi-Square | Pr > ChiSq |
| **Intercept** | 1 | 0.1844 | 0.0504 | 0.0857 | 0.2832 | 13.41 | 0.0003 |
| **LogX1** | 1 | 0.9121 | 0.0491 | 0.8158 | 1.0083 | 345.05 | <.0001 |
| **X2** | 1 | 0.0095 | 0.0004 | 0.0088 | 0.0102 | 728.62 | <.0001 |
| **X3** | 1 | 0.0096 | 0.0003 | 0.0090 | 0.0101 | 1139.73 | <.0001 |
| **Scale** | 0 | 0.0434 | 0.0000 | 0.0434 | 0.0434 | | |

**Note:** The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

**Output 51.8.8** Cumulative Residual Plot with Log(X1)



## Example 51.9: Assessment of a Marginal Model for Dependent Data

This example illustrates the use of cumulative residuals to assess the adequacy of a marginal model for dependent data fit by generalized estimating equations (GEEs). The assessment methods are applied to CD4 count data from an AIDS clinical trial reported by Fischl, Richman, and Hansen (1990) and reanalyzed by Lin, Wei, and Ying (2002). The study randomly assigned 360 HIV patients to the drug AZT and 351 patients to placebo. CD4 counts were measured repeatedly over the course of the study. The data used here are the 4328 measurements taken in the first 40 weeks of the study.

The analysis focuses on the time trend of the response. The first model considered is

$$E(y_{ik}) = \beta_0 + \beta_1 T_{ik} + \beta_2 T_{ik}^2 + \beta_3 R_i T_{ik} + \beta_4 R_i T_{ik}^2$$

where $T_{ik}$ is the time (in weeks) of the $k$th measurement on the $i$th patient, $y_{ik}$ is the CD4 count at $T_{ik}$ for the $i$th patient, and $R_i$ is the indicator of AZT for the $i$th patient. Normal errors and an independent working correlation are assumed.

The following statements create the SAS data set cd4:

```
data cd4;
   input Id Y Time Time2 TrtTime TrtTime2;
   Time3 = Time2 * Time;
   TrtTime3 = TrtTime2 * Time;
   datalines;
1     264.00024       -0.28571        0.08163       -0.28571        0.08163
1     175.00070        4.14286       17.16327        4.14286       17.16327
1     306.00150        8.14286       66.30612        8.14286       66.30612
1     331.99835       12.14286      147.44898       12.14286      147.44898
1     309.99929       16.14286      260.59184       16.14286      260.59184
1     185.00077       28.71429      824.51020       28.71429      824.51020
1     175.00070       40.14286     1611.44898       40.14286     1611.44898
2     574.99998       -0.57143        0.32653        0.00000        0.00000

   ... more lines ...

711     363.99859        8.14286       66.30612        8.14286       66.30612
711     488.00224       12.14286      147.44898       12.14286      147.44898
711     240.00026       18.14286      329.16327       18.14286      329.16327
;
```

The following SAS statements fit the preceding model, create the cumulative residual plot in Output 51.9.1, and compute a $p$-value for the model.

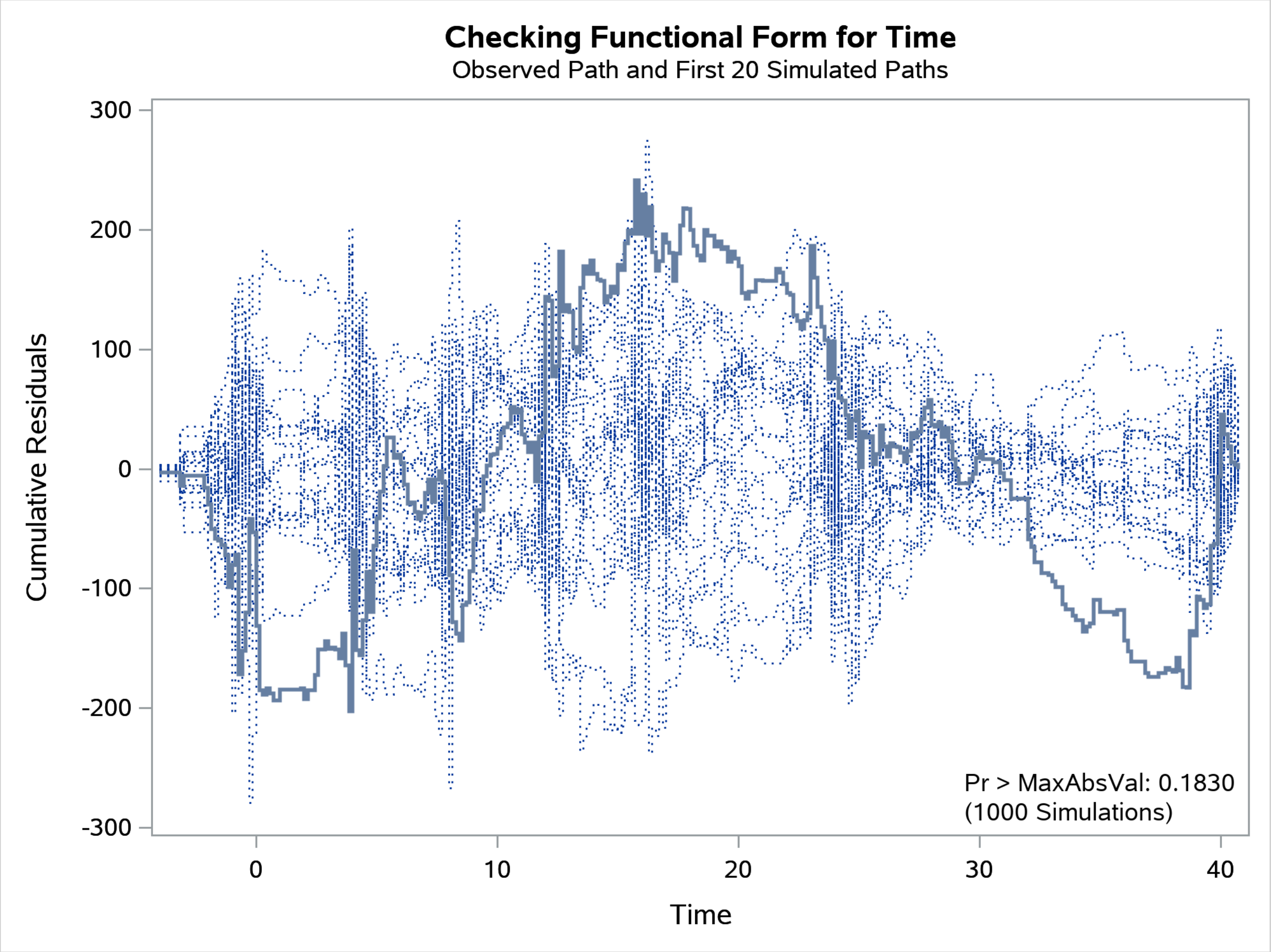
To request these graphs, ODS Graphics must be enabled and you must specify the ASSESS statement. For general information about ODS Graphics, see Chapter 24, "Statistical Graphics Using ODS." For specific information about the graphics available in the GENMOD procedure, see the section "ODS Graphics" on page 3703.

Here, the SAS data set variables Time, Time2, TrtTime, and TrtTime2 correspond to $T_{ik}$, $T_{ik}^2$, $R_i T_{ik}$, and $R_i T_{ik}^2$, respectively. The variable Id identifies individual patients.

```
ods graphics on;

proc genmod data=cd4;
   class Id;
   model Y = Time Time2 TrtTime TrtTime2;
   repeated sub=Id;
   assess var=(Time) / resample
                       seed=603708000;
run;
```

**Output 51.9.1** Cumulative Residual Plot for Quadratic Time Fit

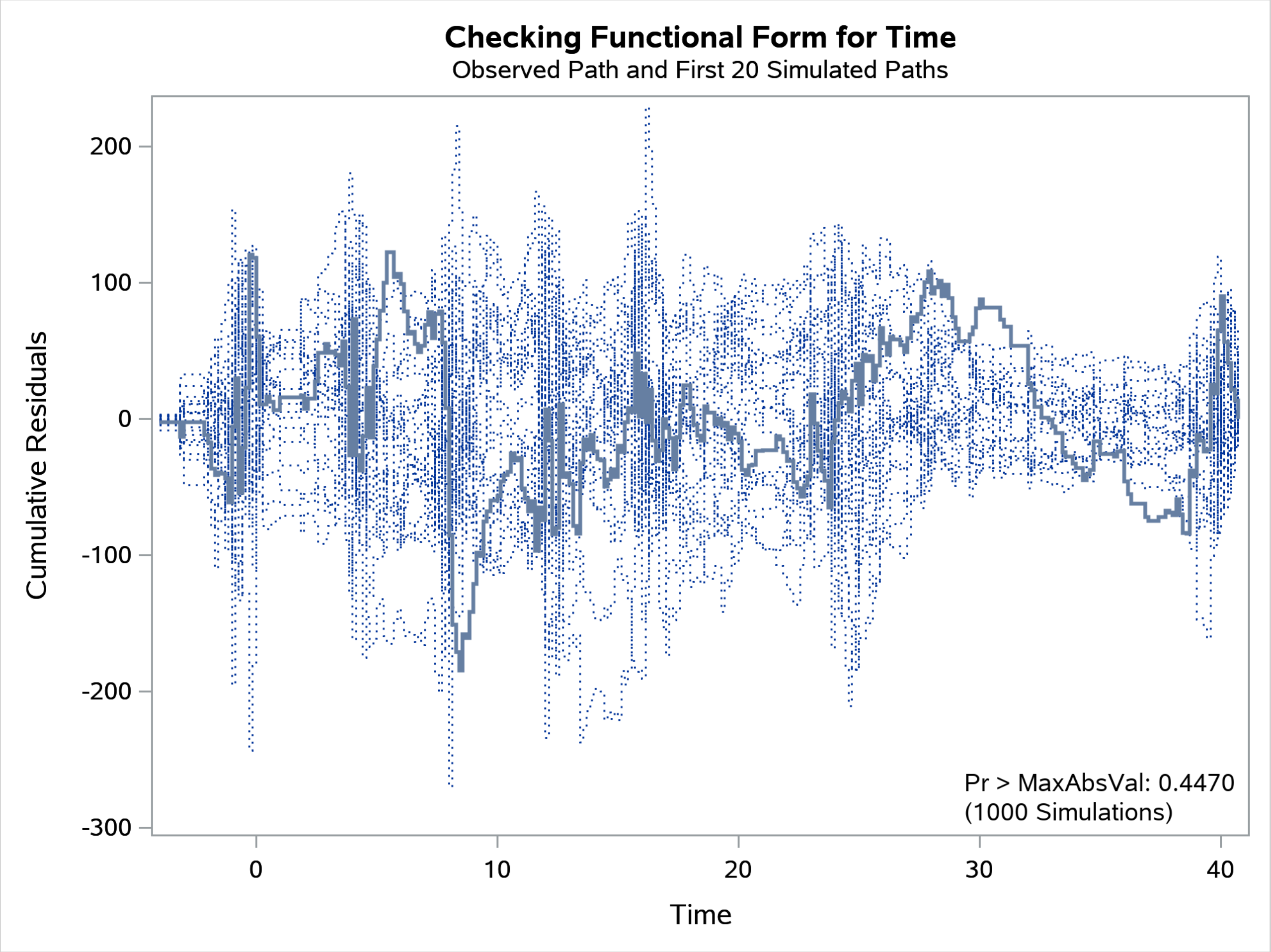


The cumulative residual plot in Output 51.9.1 displays cumulative residuals versus time for the model and 20 simulated realizations. The associated *p*-value, also shown in Output 51.9.1, is 0.18. These results indicate that a more satisfactory model might be possible. The observed cumulative residual pattern most resembles plot (c) in Output 51.8.6, suggesting cubic time trends.

The following SAS statements fit the model, create the plot in Output 51.9.2, and compute a *p*-value for a model with the additional terms $T_{ik}^3$ and $R_i T_{ik}^3$:

```
proc genmod data=cd4;
   class Id;
   model Y = Time Time2 Time3 TrtTime TrtTime2 TrtTime3;
   repeated sub=Id;
   assess var=(Time) / resample
                       seed=603708000;
run;
```

**Output 51.9.2** Cumulative Residual Plot for Cubic Time Fit



The observed cumulative residual pattern appears more typical of the simulated realizations, and the *p*-value is 0.45, indicating that the model with cubic time trends is more appropriate.

## Example 51.10: Bayesian Analysis of a Poisson Regression Model

This example illustrates a Bayesian analysis of a log-linear Poisson regression model. Consider the following data on patients from clinical trials. The data set is a subset of the data described in Ibrahim, Chen, and Lipsitz (1999).

```
data Liver;
   input X1-X6 Y;
   datalines;
19.1358    50.0110     51.000      0        0       1        3
23.5970    18.4959      3.429      0        0       1        9
20.0474    56.7699      3.429      1        1       0        6
28.0277    59.7836      4.000      0        0       1        6
28.6851    74.1589      5.714      1        0       1        1
18.8092    31.0630      2.286      0        1       1       61
```

```
28.7201    52.9178    37.286     1       0       1       6
21.3669    61.6603    54.143     0       1       1       6
23.7332    42.2904     0.571     1       0       1      21
20.4783    22.1260    19.000     1       0       1       6

   ... more lines ...

17.0993    48.8384     3.000     0       0       0      9
19.1327    65.3425     2.571     1       0       0      1
17.3010    51.4493     4.429     1       0       0      6
;
```

The primary interest is in prediction of the number of cancerous liver nodes when a patient enters the trials, by using six other baseline characteristics. The number of nodes is modeled by a Poisson regression model with the six baseline characteristics as covariates. The response and regression variables are as follows:

| | |
|---|---|
| Y | Number of Cancerous Liver Nodes |
| X1 | Body Mass Index |
| X2 | Age, in Years |
| X3 | Time Since Diagnosis of Disease, in Weeks |
| X4 | Two Biochemical Markers (each classified as normal=1 or abnormal=0) |
| X5 | Anti Hepatitis B Antigen |
| X6 | Associated Jaundice (yes=1, no=0) |

Two analyses are performed using PROC GENMOD. The first analysis uses noninformative normal prior distributions, and the second analysis uses an informative normal prior for one of the regression parameters.

In the following BAYES statement, COEFFPRIOR=NORMAL specifies a noninformative independent normal prior distribution with zero mean and variance $10^6$ for each parameter.

The initial analysis is performed using PROC GENMOD to obtain Bayesian estimates of the regression coefficients by using the following SAS statements:

```
proc genmod data=Liver;
   model Y = X1-X6 / dist=Poisson link=log;
   bayes seed=1 coeffprior=normal;
run;
```

Maximum likelihood estimates of the model parameters are computed by default. These are shown in the "Analysis of Maximum Likelihood Parameter Estimates" table in Output 51.10.1.

**Output 51.10.1** Maximum Likelihood Parameter Estimates

**The GENMOD Procedure**

**Bayesian Analysis**

**Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | |
|---|---|---|---|---|---|
| **Intercept** | 1 | 2.4508 | 0.2284 | 2.0032 | 2.8984 |
| **X1** | 1 | -0.0044 | 0.0080 | -0.0201 | 0.0114 |
| **X2** | 1 | -0.0135 | 0.0024 | -0.0181 | -0.0088 |
| **X3** | 1 | -0.0029 | 0.0022 | -0.0072 | 0.0014 |
| **X4** | 1 | -0.2715 | 0.0795 | -0.4272 | -0.1157 |
| **X5** | 1 | 0.3215 | 0.0832 | 0.1585 | 0.4845 |
| **X6** | 1 | 0.2077 | 0.0827 | 0.0456 | 0.3698 |
| **Scale** | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 |

**Note:** The scale parameter was held fixed.

Noninformative independent normal prior distributions with zero means and variances of $10^6$ were used in the initial analysis. These are shown in Output 51.10.2.

**Output 51.10.2** Regression Coefficient Priors

**The GENMOD Procedure**

**Bayesian Analysis**

**Independent Normal Prior for Regression Coefficients**

| Parameter | Mean | Precision |
|---|---|---|
| Intercept | 0 | 1E-6 |
| X1 | 0 | 1E-6 |
| X2 | 0 | 1E-6 |
| X3 | 0 | 1E-6 |
| X4 | 0 | 1E-6 |
| X5 | 0 | 1E-6 |
| X6 | 0 | 1E-6 |

Initial values for the Markov chain are listed in the “Initial Values and Seeds” table in Output 51.10.3. The random number seed is also listed so that you can reproduce the analysis. Since no seed was specified, the seed shown was derived from the time of day.

**Output 51.10.3** MCMC Initial Values and Seeds

**Initial Values of the Chain**

| Chain | Seed | Intercept | X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2.450813 | -0.00435 | -0.01347 | -0.00291 | -0.27149 | 0.321507 | 0.207713 |

Summary statistics for the posterior sample are displayed in the "Fit Statistics," "Descriptive Statistics for the Posterior Sample," "Interval Statistics for the Posterior Sample," and "Posterior Correlation Matrix" tables in Output 51.10.4, Output 51.10.5, Output 51.10.6, and Output 51.10.7, respectively. Since noninformative prior distributions for the regression coefficients were used, the mean and standard deviations of the posterior distributions for the model parameters are close to the maximum likelihood estimates and standard errors.

**Output 51.10.4** Fit Statistics

| Fit Statistics | |
|---|---|
| DIC (smaller is better) | 829.810 |
| pD (effective number of parameters) | 7.005 |

**Output 51.10.5** Descriptive Statistics

**The GENMOD Procedure**

**Bayesian Analysis**

Posterior Summaries

| | | | | Percentiles | | |
|---|---|---|---|---|---|---|
| Parameter | N | Mean | Standard Deviation | 25% | 50% | 75% |
| **Intercept** | 10000 | 2.4483 | 0.2320 | 2.2903 | 2.4493 | 2.6093 |
| **X1** | 10000 | -0.00475 | 0.00809 | -0.0101 | -0.00466 | 0.000851 |
| **X2** | 10000 | -0.0134 | 0.00237 | -0.0150 | -0.0134 | -0.0118 |
| **X3** | 10000 | -0.00303 | 0.00220 | -0.00445 | -0.00298 | -0.00150 |
| **X4** | 10000 | -0.2703 | 0.0799 | -0.3241 | -0.2725 | -0.2190 |
| **X5** | 10000 | 0.3202 | 0.0828 | 0.2642 | 0.3209 | 0.3775 |
| **X6** | 10000 | 0.2106 | 0.0838 | 0.1533 | 0.2111 | 0.2663 |

**Output 51.10.6** Interval Statistics

Posterior Intervals

| Parameter | Alpha | Equal-Tail Interval | | HPD Interval | |
|---|---|---|---|---|---|
| **Intercept** | 0.050 | 1.9903 | 2.9059 | 2.0289 | 2.9321 |
| **X1** | 0.050 | -0.0209 | 0.0108 | -0.0211 | 0.0106 |
| **X2** | 0.050 | -0.0181 | -0.00870 | -0.0184 | -0.00908 |
| **X3** | 0.050 | -0.00761 | 0.00105 | -0.00745 | 0.00113 |
| **X4** | 0.050 | -0.4257 | -0.1063 | -0.4314 | -0.1152 |
| **X5** | 0.050 | 0.1563 | 0.4804 | 0.1574 | 0.4811 |
| **X6** | 0.050 | 0.0450 | 0.3777 | 0.0468 | 0.3788 |

**Output 51.10.7** Posterior Sample Correlation Matrix

**Posterior Correlation Matrix**

| Parameter | Intercept | X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|---|---|
| **Intercept** | 1.000 | -0.708 | -0.432 | -0.046 | -0.261 | -0.185 | -0.422 |
| **X1** | -0.708 | 1.000 | -0.202 | -0.047 | -0.035 | 0.078 | 0.129 |
| **X2** | -0.432 | -0.202 | 1.000 | 0.035 | 0.076 | 0.054 | 0.117 |
| **X3** | -0.046 | -0.047 | 0.035 | 1.000 | 0.027 | -0.042 | -0.077 |
| **X4** | -0.261 | -0.035 | 0.076 | 0.027 | 1.000 | -0.024 | 0.127 |
| **X5** | -0.185 | 0.078 | 0.054 | -0.042 | -0.024 | 1.000 | -0.037 |
| **X6** | -0.422 | 0.129 | 0.117 | -0.077 | 0.127 | -0.037 | 1.000 |

Posterior sample autocorrelations for each model parameter are shown in Output 51.10.8. The autocorrelation after 10 lags is negligible for all parameters, indicating good mixing in the Markov chain.

**Output 51.10.8** Posterior Sample Autocorrelations

**The GENMOD Procedure**

**Bayesian Analysis**

**Posterior Autocorrelations**

| Parameter | Lag 1 | Lag 5 | Lag 10 | Lag 50 |
|---|---|---|---|---|
| **Intercept** | 0.3037 | 0.0152 | 0.0095 | -0.0170 |
| **X1** | 0.3398 | 0.0025 | 0.0003 | 0.0052 |
| **X2** | 0.3036 | 0.0061 | 0.0003 | -0.0062 |
| **X3** | 0.3489 | 0.0190 | -0.0064 | -0.0210 |
| **X4** | 0.2868 | 0.0213 | 0.0157 | -0.0107 |
| **X5** | 0.2854 | 0.0108 | -0.0288 | -0.0012 |
| **X6** | 0.3078 | 0.0230 | 0.0073 | 0.0062 |

The *p*-values for the Geweke test statistics shown in Output 51.10.9 all indicate convergence of the MCMC. See the section "Assessing Markov Chain Convergence" on page 162 in Chapter 8, "Introduction to Bayesian Analysis Procedures," for more information about convergence diagnostics and their interpretation.

**Output 51.10.9** Geweke Diagnostic Statistics

**Geweke Diagnostics**

| Parameter | z | Pr > \|z\| |
|---|---|---|
| **Intercept** | -0.6533 | 0.5135 |
| **X1** | 0.3418 | 0.7325 |
| **X2** | 0.3609 | 0.7182 |
| **X3** | -0.3345 | 0.7380 |
| **X4** | 0.2851 | 0.7755 |
| **X5** | -0.5266 | 0.5985 |
| **X6** | 1.1285 | 0.2591 |

The effective sample sizes for each parameter are shown in Output 51.10.10.

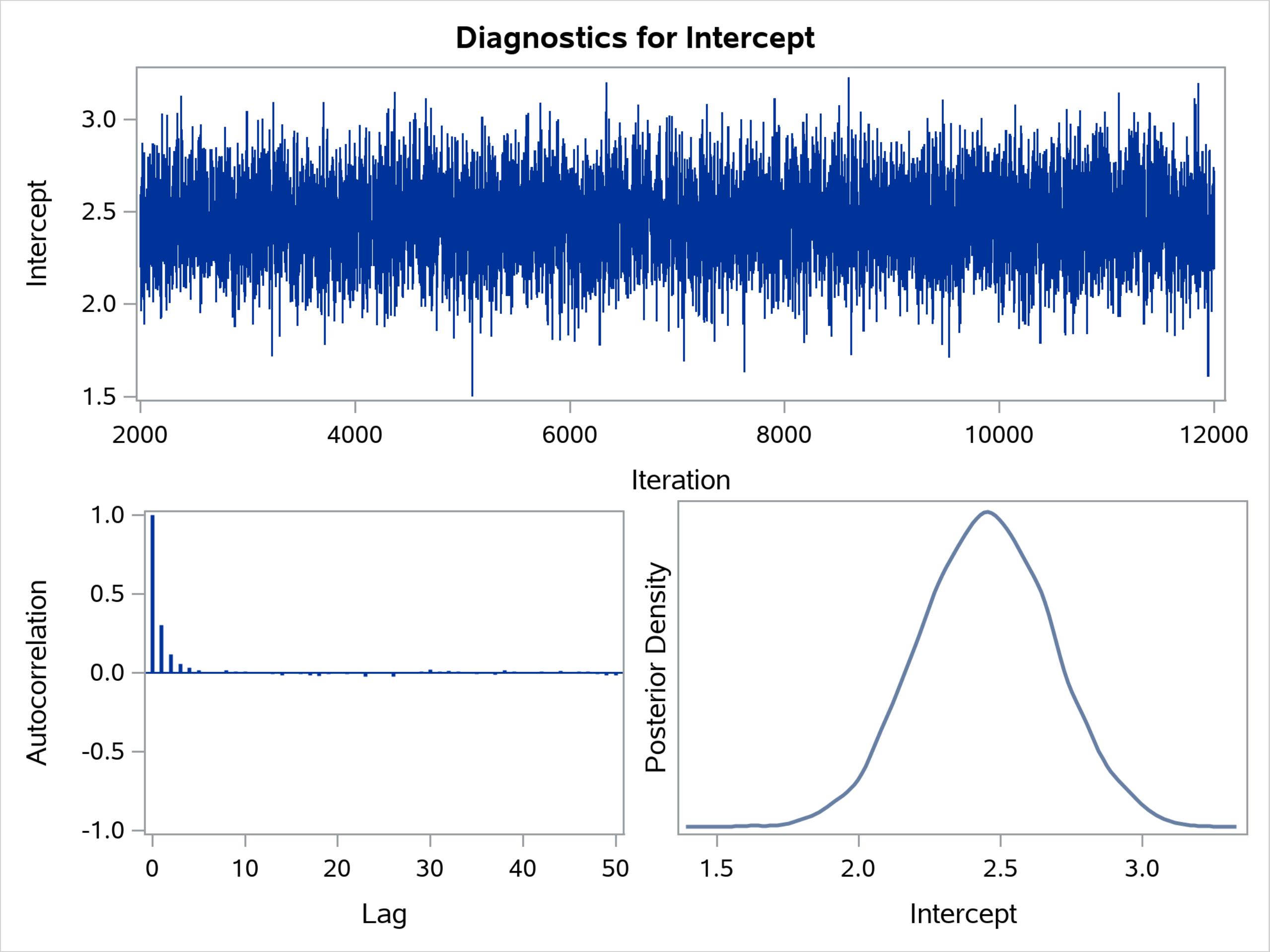
**Output 51.10.10** Effective Sample Sizes

**Effective Sample Sizes**

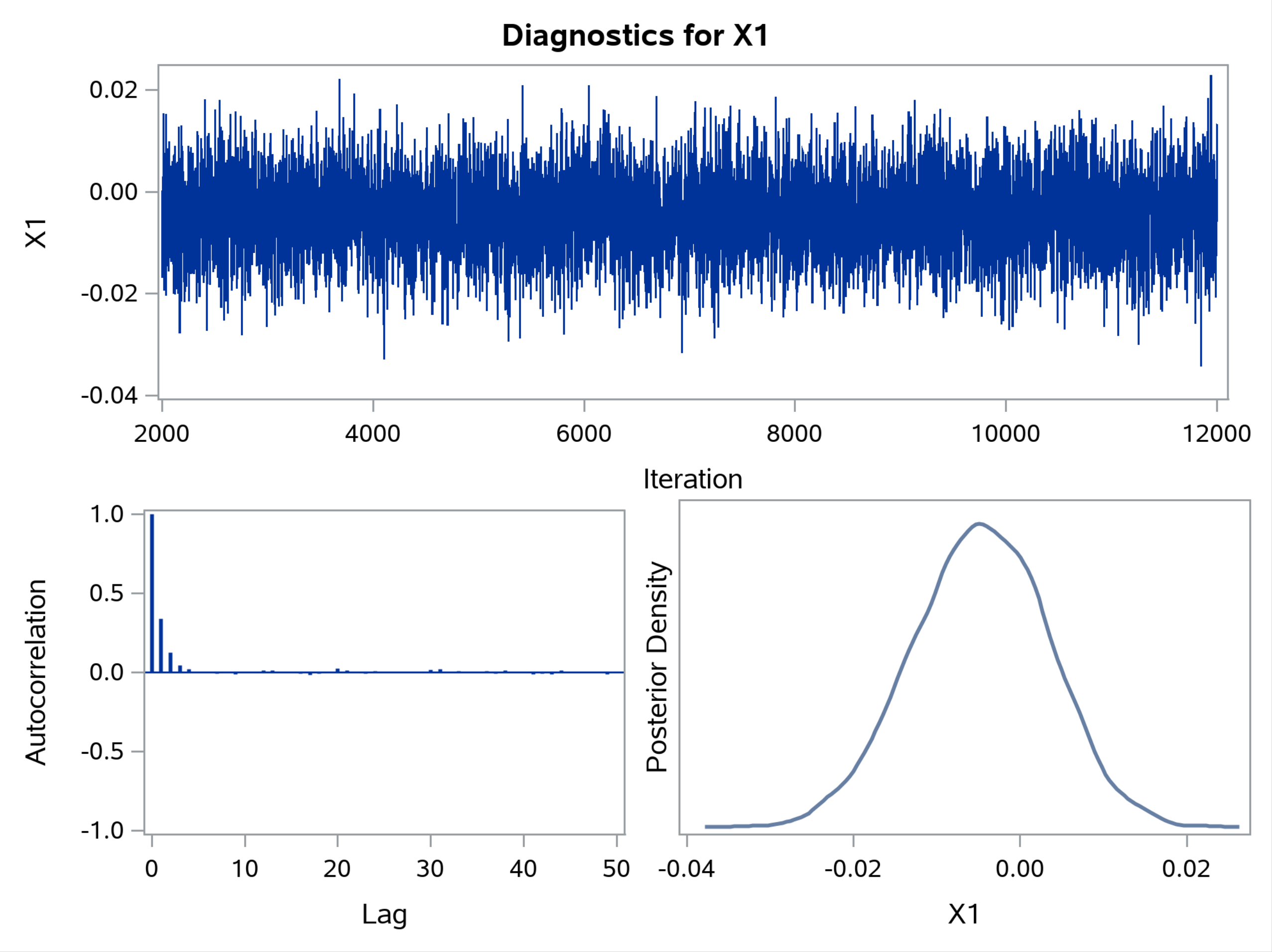
| Parameter | ESS | Autocorrelation Time | Efficiency |
|---|---|---|---|
| **Intercept** | 4880.3 | 2.0491 | 0.4880 |
| **X1** | 4844.2 | 2.0643 | 0.4844 |
| **X2** | 5139.3 | 1.9458 | 0.5139 |
| **X3** | 4551.2 | 2.1972 | 0.4551 |
| **X4** | 4953.6 | 2.0187 | 0.4954 |
| **X5** | 5330.5 | 1.8760 | 0.5331 |
| **X6** | 4988.1 | 2.0048 | 0.4988 |

Trace, autocorrelation, and density plots for the seven model parameters are shown in Output 51.10.11 through Output 51.10.17. All indicate satisfactory convergence of the Markov chain.

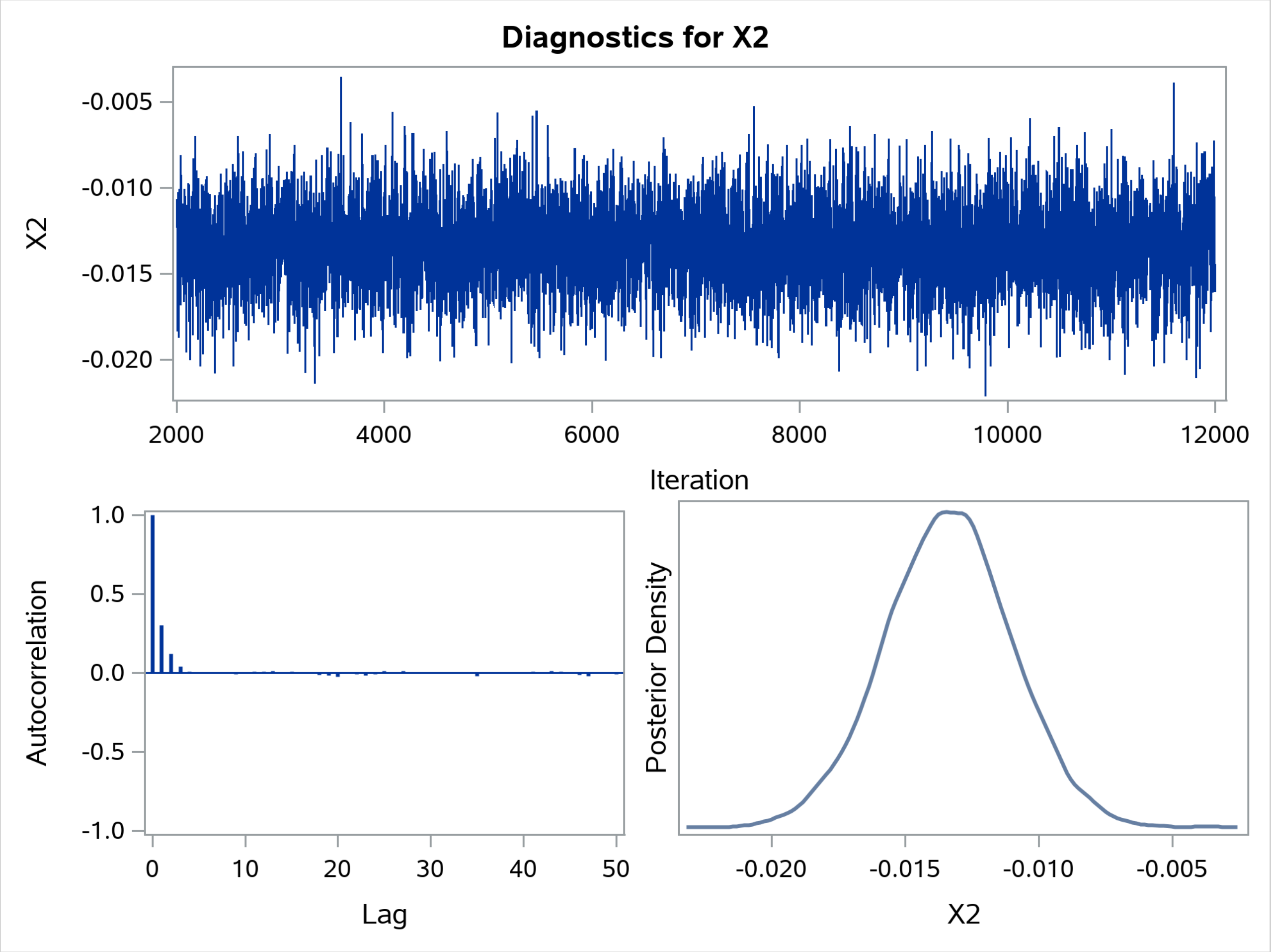
**Output 51.10.11** Diagnostic Plots for Intercept



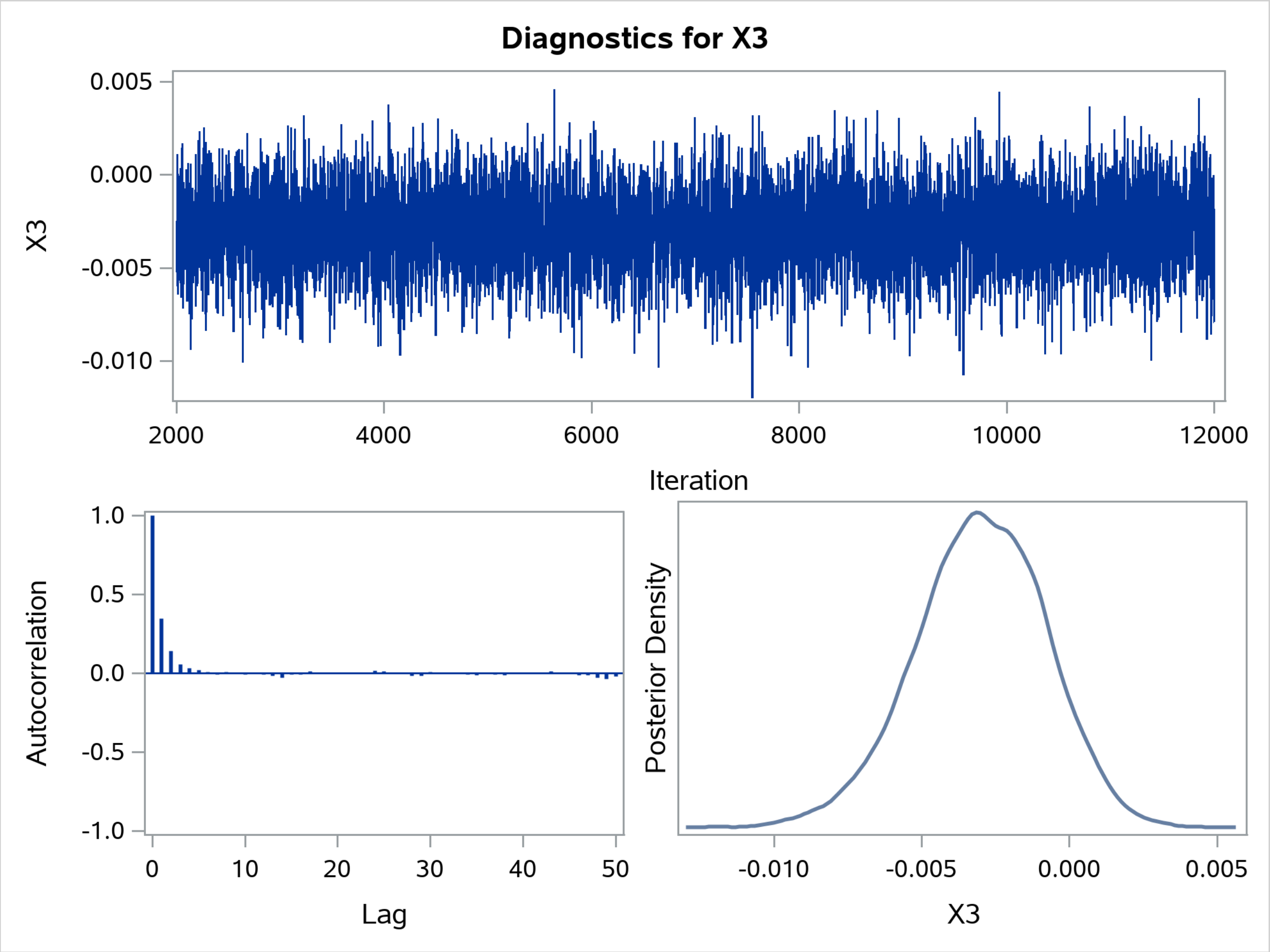
**Output 51.10.12** Diagnostic Plots for X1



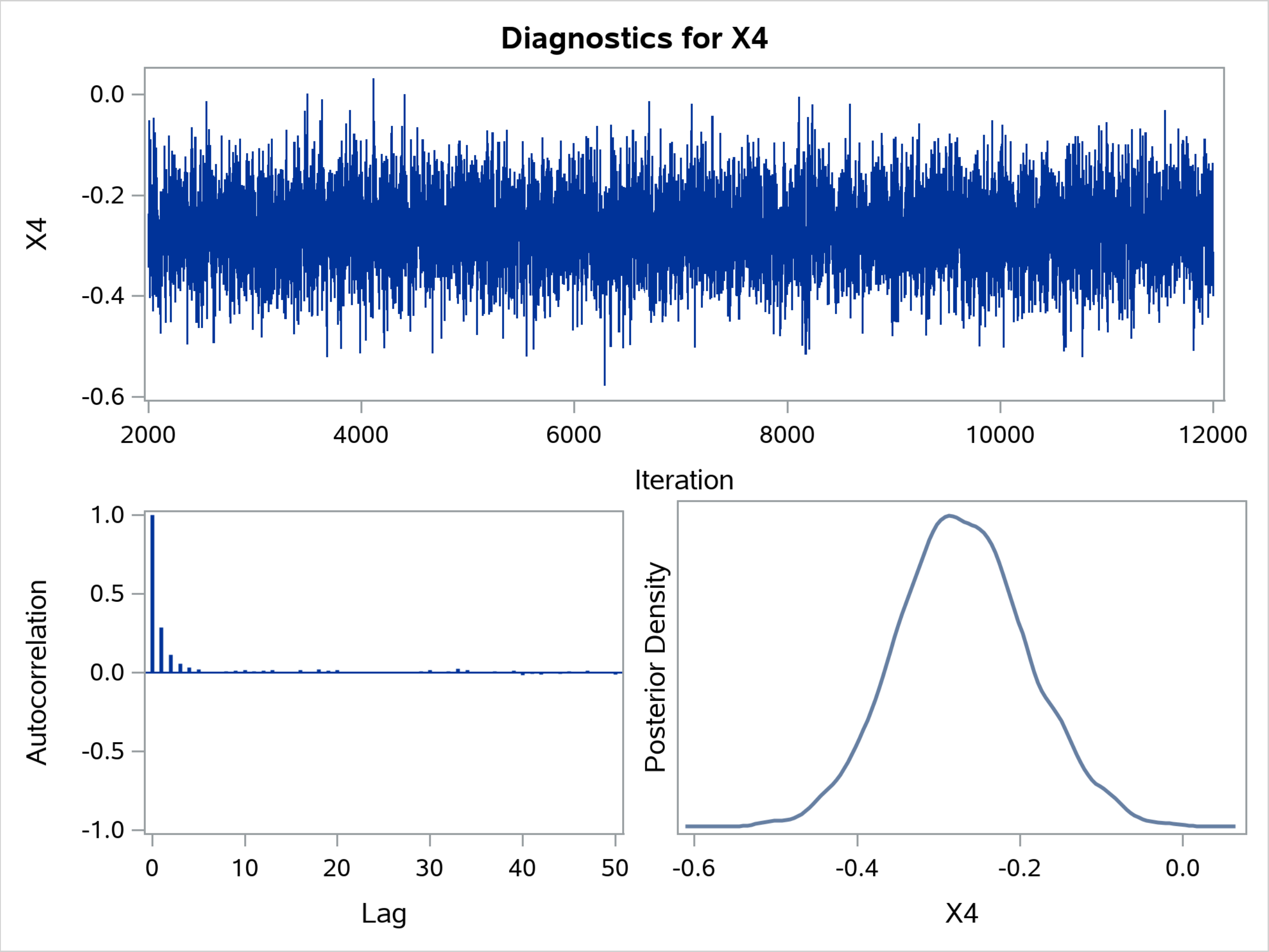
**Output 51.10.13** Diagnostic Plots for X2



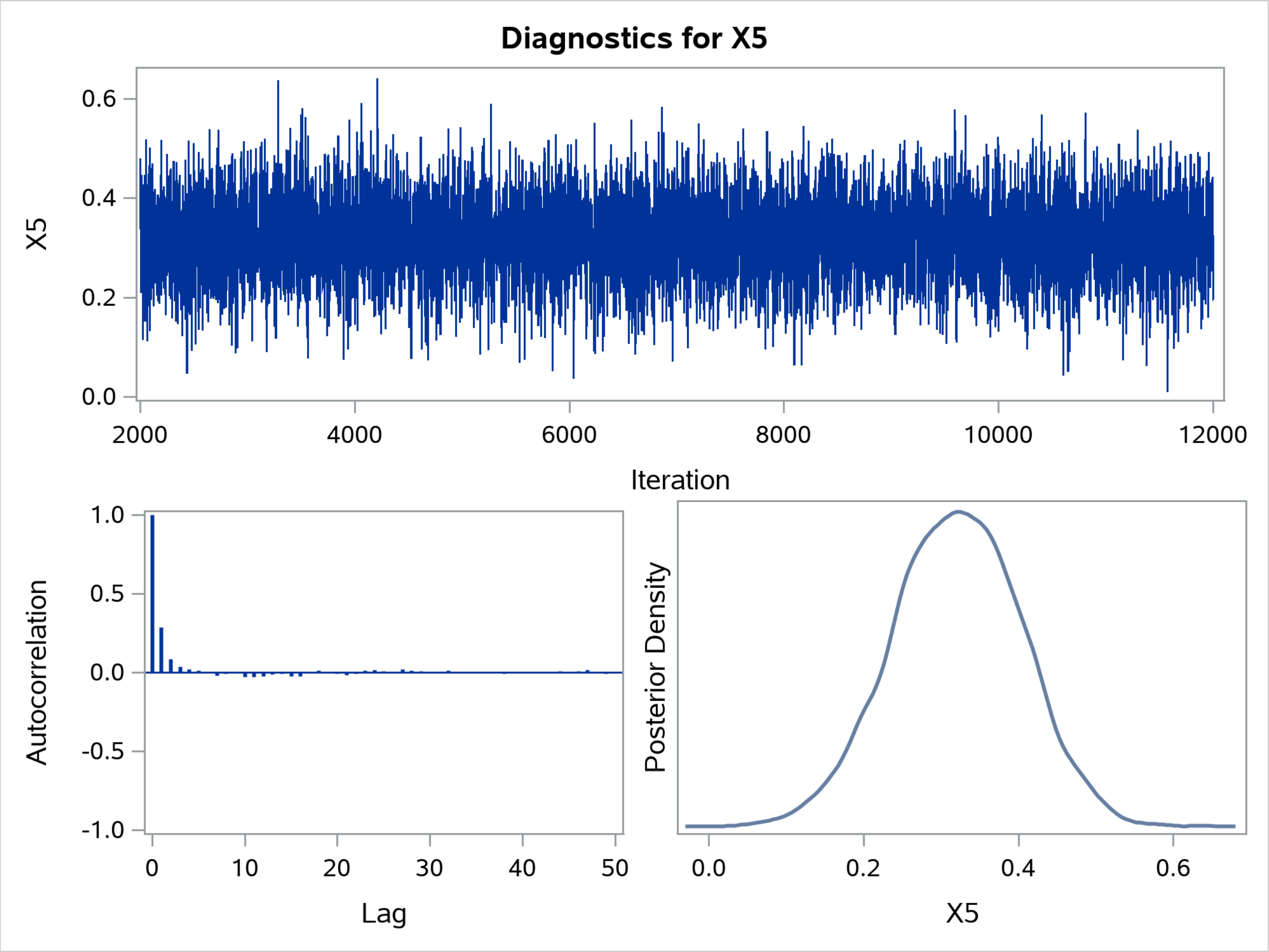
**Output 51.10.14** Diagnostic Plots for X3



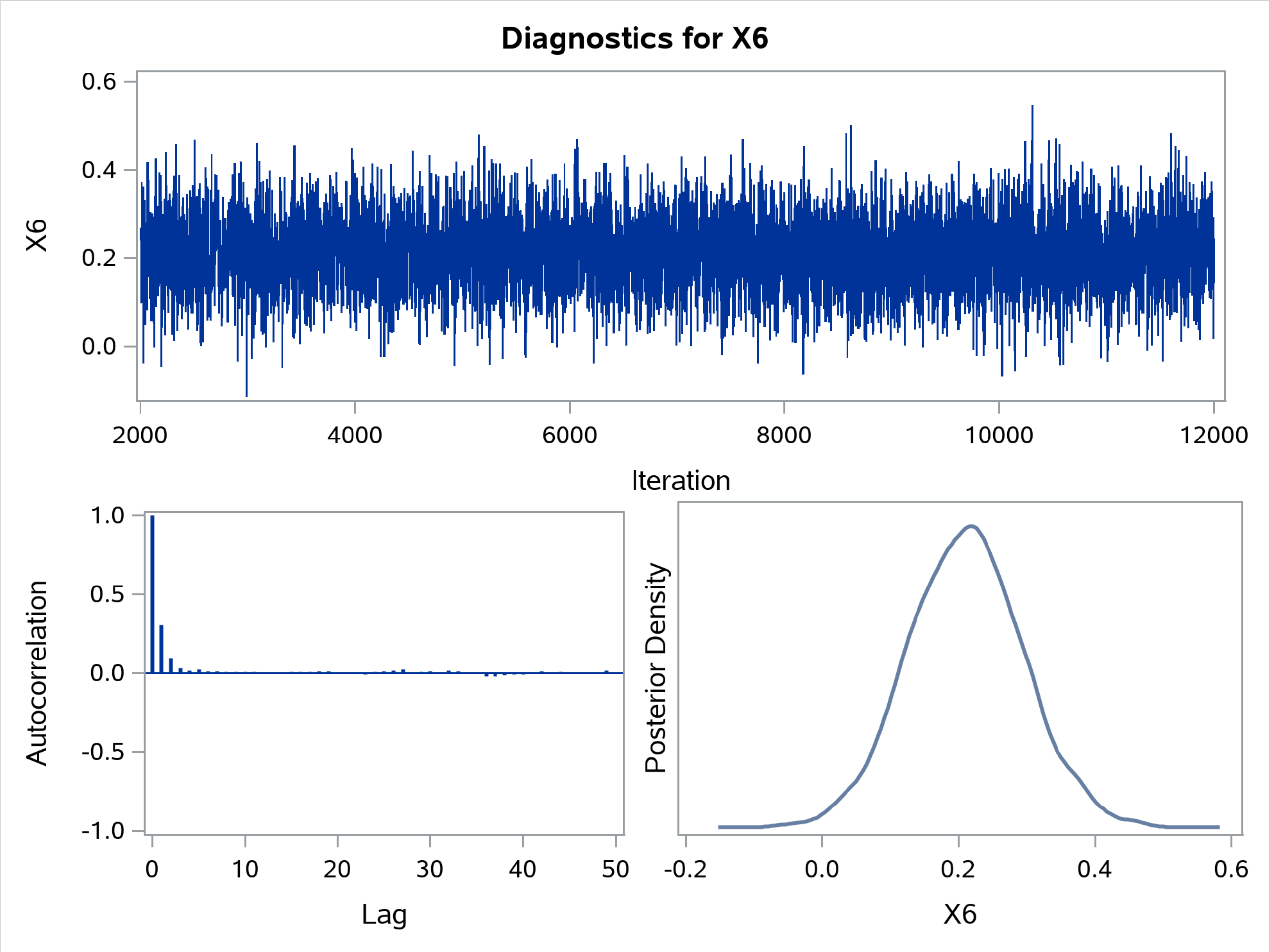
**Output 51.10.15** Diagnostic Plots for X4



**Output 51.10.16** Diagnostic Plots for X5



**Output 51.10.17** Diagnostic Plots for X6



In order to illustrate the use of an informative prior distribution, suppose that researchers expect that a unit increase in body mass index (X1) will be associated with an increase in the mean number of nodes of between 10% and 20%, and they want to incorporate this prior knowledge in the Bayesian analysis. For log-linear models, the mean and linear predictor are related by $\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$. If $\mathsf{X1}_1$ and $\mathsf{X1}_2$ are two values of body mass index, $\mu_1$ and $\mu_2$ are the two mean values, and all other covariates remain equal for the two values of X1, then

$$\frac{\mu_1}{\mu_2} = \exp(\beta(\mathsf{X1}_1 - \mathsf{X1}_2))$$

so that for a unit change in X1,

$$\frac{\mu_1}{\mu_2} = \exp(\beta)$$

If $1.1 \le \frac{\mu_1}{\mu_2} \le 1.2$, then $1.1 \le \exp(\beta) \le 1.2$, or $0.095 \le \beta \le 0.182$. This gives you guidance in specifying a prior distribution for the $\beta$ for body mass index. Taking the mean of the prior normal distribution to be the midrange of the values of $\beta$, and taking $\mu \pm 2\sigma$ to be the extremes of the range, an $N(0.1385, 0.0005)$ is the resulting prior distribution. The second analysis uses this informative normal prior distribution for the coefficient of X1 and uses independent noninformative normal priors with zero means and variances equal to $10^6$ for the remaining model regression parameters.

In the following BAYES statement, COEFFPRIOR=NORMAL(INPUT=NormalPrior) specifies the normal prior distribution for the regression coefficients with means and variances contained in the data set NormalPrior.

An analysis is performed using PROC GENMOD to obtain Bayesian estimates of the regression coefficients by using the following SAS statements:

```
data NormalPrior;
   input _type_ $ Intercept X1-X6;
   datalines;
Var  1e6   0.0005      1e6      1e6      1e6      1e6      1e6
Mean 0.0   0.1385      0.0      0.0      0.0      0.0      0.0
;

proc genmod data=Liver;
   model Y = X1-X6 / dist=Poisson link=log;
   bayes seed=1 plots=none coeffprior=normal(input=NormalPrior);
run;
```

The prior distributions for the regression parameters are shown in Output 51.10.18.

**Output 51.10.18** Regression Coefficient Priors

**The GENMOD Procedure**

**Bayesian Analysis**

**Independent Normal Prior for Regression Coefficients**

| Parameter | Mean | Precision |
|---|---|---|
| Intercept | 0 | 1E-6 |
| X1 | 0.1385 | 2000 |
| X2 | 0 | 1E-6 |
| X3 | 0 | 1E-6 |
| X4 | 0 | 1E-6 |
| X5 | 0 | 1E-6 |
| X6 | 0 | 1E-6 |

Initial values for the MCMC are shown in Output 51.10.19. The initial values of the covariates are joint estimates of their posterior modes. The prior distribution for X1 is informative, so the initial value of X1 is further from the MLE than the rest of the covariates. Initial values for the rest of the covariates are close to their MLEs, since noninformative prior distributions were specified for them.

**Output 51.10.19** MCMC Initial Values and Seeds

**Initial Values of the Chain**

| Chain | Seed | Intercept | X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2.14282 | 0.010595 | -0.01434 | -0.00301 | -0.28062 | 0.334983 | 0.231213 |

Goodness-of-fit, summary, and interval statistics are shown in Output 51.10.20. Except for X1, the statistics shown in Output 51.10.20 are very similar to the previous statistics for noninformative priors shown in Output 51.10.4 through Output 51.10.7. The point estimate for X1 is now positive. This is expected because the prior distribution on $\beta_1$ is quite informative. The distribution reflects the belief that the coefficient is positive. The $N(0.1385, 0.0005)$ distribution places the majority of its probability density on positive values. As a result, the posterior density of $\beta_1$ places more likelihood on positive values than in the noninformative case.

**Output 51.10.20** Fit Statistics

| Fit Statistics | |
|---|---|
| DIC (smaller is better) | 833.074 |
| pD (effective number of parameters) | 6.869 |

**The GENMOD Procedure**

**Bayesian Analysis**

**Posterior Summaries**

| | | | | Percentiles | | |
|---|---|---|---|---|---|---|
| | | | Standard | | | |
| **Parameter** | **N** | **Mean** | **Deviation** | **25%** | **50%** | **75%** |
| **Intercept** | 10000 | 2.1419 | 0.2157 | 1.9965 | 2.1430 | 2.2894 |
| **X1** | 10000 | 0.0103 | 0.00684 | 0.00573 | 0.0104 | 0.0150 |
| **X2** | 10000 | -0.0143 | 0.00233 | -0.0159 | -0.0142 | -0.0127 |
| **X3** | 10000 | -0.00318 | 0.00218 | -0.00467 | -0.00314 | -0.00170 |
| **X4** | 10000 | -0.2806 | 0.0800 | -0.3336 | -0.2793 | -0.2266 |
| **X5** | 10000 | 0.3341 | 0.0832 | 0.2788 | 0.3341 | 0.3906 |
| **X6** | 10000 | 0.2333 | 0.0826 | 0.1774 | 0.2325 | 0.2880 |

**Posterior Intervals**

| | | Equal-Tail | | | |
|---|---|---|---|---|---|
| **Parameter** | **Alpha** | **Interval** | | **HPD Interval** | |
| **Intercept** | 0.050 | 1.7225 | 2.5574 | 1.7293 | 2.5632 |
| **X1** | 0.050 | -0.00344 | 0.0235 | -0.00345 | 0.0234 |
| **X2** | 0.050 | -0.0188 | -0.00970 | -0.0189 | -0.00980 |
| **X3** | 0.050 | -0.00757 | 0.00108 | -0.00733 | 0.00121 |
| **X4** | 0.050 | -0.4365 | -0.1200 | -0.4391 | -0.1256 |
| **X5** | 0.050 | 0.1657 | 0.4966 | 0.1682 | 0.4987 |
| **X6** | 0.050 | 0.0695 | 0.3959 | 0.0725 | 0.3981 |

## Example 51.11: Exact Poisson Regression

The following data, taken from Cox and Snell (1989, pp. 10–11), consists of the number, Notready, of ingots that are not ready for rolling, out of Total tested, for several combinations of heating time and soaking time:

```
data ingots;
   input Heat Soak Notready Total @@;
   lnTotal= log(Total);
   datalines;
7 1.0 0 10   14 1.0 0 31   27 1.0 1 56   51 1.0 3 13
7 1.7 0 17   14 1.7 0 43   27 1.7 4 44   51 1.7 0  1
7 2.2 0  7   14 2.2 2 33   27 2.2 0 21   51 2.2 0  1
7 2.8 0 12   14 2.8 0 31   27 2.8 1 22   51 4.0 0  1
7 4.0 0  9   14 4.0 0 19   27 4.0 1 16
;
```

The following invocation of PROC GENMOD fits an asymptotic (unconditional) Poisson regression model to the data. The variable Notready is specified as the response variable, and the continuous predictors Heat and Soak are defined in the CLASS statement as categorical predictors that use reference coding. Specifying the offset variable as lnTotal enables you to model the ratio Notready/Total.

```
proc genmod data=ingots;
   class Heat Soak / param=ref;
   model Notready=Heat Soak / offset=lnTotal dist=Poisson link=log;
   exact Heat Soak / joint estimate;
   exactoptions statustime=10;
run;
```

The EXACT statement is specified to additionally fit an exact conditional Poisson regression model. Specifying the lnTotal offset variable models the ratio Notready/Total; in this case, the Total variable contains the largest possible response value for each observation. The JOINT option produces a joint test for the significance of the covariates, along with the usual marginal tests. The ESTIMATE option produces exact parameter estimates for the covariates. The STATUSTIME=10 option is specified in the EXACTOPTIONS statement for monitoring the progress of the results; this example can take several minutes to complete due to the JOINT option. If you run out of memory, see the SAS Companion for your system for information about how to increase the available memory.

The “Criteria For Assessing Goodness Of Fit” table is displayed in Output 51.11.1. Comparing the deviance of 10.9363 to an asymptotic chi-square distribution with 11 degrees of freedom, you find that the *p*-value is 0.449. This indicates that the specified model fits the data reasonably well.

**Output 51.11.1** Unconditional Goodness of Fit Criteria

**The GENMOD Procedure**

| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| **Criterion** | **DF** | **Value** | **Value/DF** |
| **Deviance** | 11 | 10.9363 | 0.9942 |
| **Scaled Deviance** | 11 | 10.9363 | 0.9942 |
| **Pearson Chi-Square** | 11 | 9.3722 | 0.8520 |
| **Scaled Pearson X2** | 11 | 9.3722 | 0.8520 |
| **Log Likelihood** | | -7.2408 | |
| **Full Log Likelihood** | | -12.9038 | |
| **AIC (smaller is better)** | | 41.8076 | |
| **AICC (smaller is better)** | | 56.2076 | |
| **BIC (smaller is better)** | | 49.3631 | |

From the "Analysis Of Parameter Estimates" table in Output 51.11.2, you can see that only two of the Heat parameters are deemed significant. Looking at the standard errors, you can see that the unconditional analysis had convergence difficulties with the Heat=7 parameter (Standard Error=264324.6), which means you cannot fit this unconditional Poisson regression model to this data.

**Output 51.11.2** Unconditional Maximum Likelihood Parameter Estimates

| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Wald 95% | | | |
| | | | | Standard | Confidence | | Wald | |
| **Parameter** | | **DF** | **Estimate** | **Error** | **Limits** | | **Chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -1.5700 | 1.1657 | -3.8548 | 0.7147 | 1.81 | 0.1780 |
| **Heat** | 7 | 1 | -27.6129 | 264324.6 | -518094 | 518039.0 | 0.00 | 0.9999 |
| **Heat** | 14 | 1 | -3.0107 | 1.0025 | -4.9756 | -1.0458 | 9.02 | 0.0027 |
| **Heat** | 27 | 1 | -1.7180 | 0.7691 | -3.2253 | -0.2106 | 4.99 | 0.0255 |
| **Soak** | 1 | 1 | -0.2454 | 1.1455 | -2.4906 | 1.9998 | 0.05 | 0.8304 |
| **Soak** | 1.7 | 1 | 0.5572 | 1.1217 | -1.6412 | 2.7557 | 0.25 | 0.6193 |
| **Soak** | 2.2 | 1 | 0.4079 | 1.2260 | -1.9951 | 2.8109 | 0.11 | 0.7394 |
| **Soak** | 2.8 | 1 | -0.1301 | 1.4234 | -2.9199 | 2.6597 | 0.01 | 0.9272 |
| **Scale** | | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

**Note:** The scale parameter was held fixed.

Following the output from the asymptotic analysis, the exact conditional Poisson regression results are displayed, as shown in Output 51.11.3.

**Output 51.11.3** Exact Tests

**The GENMOD Procedure**

**Exact Conditional Analysis**

**Exact Conditional Tests**

| | | | p-Value | |
|---|---|---|---|---|
| **Effect** | **Test** | **Statistic** | **Exact** | **Mid** |
| **Joint** | **Score** | 18.3665 | 0.0137 | 0.0137 |
| | **Probability** | 1.294E-6 | 0.0471 | 0.0471 |
| **Heat** | **Score** | 15.8259 | 0.0023 | 0.0022 |
| | **Probability** | 0.000175 | 0.0063 | 0.0062 |
| **Soak** | **Score** | 1.4612 | 0.8683 | 0.8646 |
| | **Probability** | 0.00735 | 0.8176 | 0.8139 |

The Joint test in the "Conditional Exact Tests" table in Output 51.11.3 is produced by specifying the JOINT option in the EXACT statement. The *p*-values for this test indicate that the parameters for Heat and Soak are jointly significant as explanatory effects in the model. If the Heat variable is the only explanatory variable in your model, then the rows of this table labeled as "Heat" show the joint significance of all the Heat effect parameters in that reduced model. In this case, a model that contains only the Heat parameters still explains a significant amount of the variability; however, you can see that a model that contains only the Soak parameters would not be significant.

The "Exact Parameter Estimates" table in Output 51.11.4 displays parameter estimates and tests of significance for the levels of the CLASS variables. Again, the Heat=7 parameter has some difficulties; however, in the exact analysis, a *median unbiased estimate* is computed for the parameter instead of a maximum likelihood estimate. The confidence limits show that the Heat variable contains some explanatory power, while the categorical Soak variable is insignificant and can be dropped from the model.

**Output 51.11.4** Exact Parameter Estimates

**Exact Parameter Estimates**

| **Parameter** | | **Estimate** | **Standard Error** | **95% Confidence Limits** | | **p-Value** |
|---|---|---|---|---|---|---|
| **Heat** | **7** | -2.7552 * | . | -Infinity | -0.7864 | 0.0099 |
| **Heat** | **14** | -3.0255 | 1.0128 | -5.7450 | -0.6194 | 0.0113 |
| **Heat** | **27** | -1.7846 | 0.8065 | -3.6779 | 0.2260 | 0.0844 |
| **Soak** | **1** | -0.3231 | 1.1717 | -2.8673 | 3.6754 | 1.0000 |
| **Soak** | **1.7** | 0.5375 | 1.1284 | -1.8056 | 4.4588 | 1.0000 |
| **Soak** | **2.2** | 0.4035 | 1.2347 | -2.5785 | 4.5054 | 1.0000 |
| **Soak** | **2.8** | -0.1661 | 1.4214 | -4.5490 | 4.2168 | 1.0000 |

**Note:** * indicates a median unbiased estimate and a one-sided p-value.

**NOTE:** If you want to make predictions from the exact results, you can obtain an estimate for the intercept parameter by specifying the INTERCEPT keyword in the EXACT statement. You should also remove the JOINT option to reduce the amount of time and memory consumed.

## Example 51.12: Tweedie Regression

The following SAS statements simulate 250 observations, which are based on an underlying Tweedie generalized linear model (GLM) that exploits its connection with the compound Poisson distribution. A natural logarithm link function is assumed for modeling the response variable (yTweedie), and there are five categorical variables (C1–C5), each of which has four numerical levels and two continuous variables (D1 and D2). By design, two of the categorical variables, C3 and C4, and one of the two continuous variables, D2, have no effect on the response. The dispersion parameter is set to 0.5, and the power parameter is set to 1.5.

```
%let nObs = 250;
%let nClass = 5;
%let nLevs = 4;
%let seed = 100;

data tmp1;
   array c{&nClass};

   keep c1-c&nClass yTweedie d1 d2;

   /* Tweedie parms */
   phi=0.5;
   p=1.5;

   do i=1 to &nObs;

      do j=1 to &nClass;
         c{j} = int(ranuni(1)*&nLevs);
      end;

      d1 = ranuni(&seed);
      d2 = ranuni(&seed);

      xBeta   =  0.5*((c2<2) - 2*(c1=1) + 0.5*c&nClass + 0.05*d1);
      mu      =  exp(xBeta);

      /* Poisson distributions parms */
      lambda = mu**(2-p)/(phi*(2-p));
      /* Gamma distribution parms */
      alpha = (2-p)/(p-1);
      gamma = phi*(p-1)*(mu**(p-1));

      rpoi = ranpoi(&seed,lambda);
      if rpoi=0 then yTweedie=0;
      else do;
         yTweedie=0;
         do j=1 to rpoi;
         yTweedie = yTweedie + rangam(&seed,alpha);
         end;
         yTweedie = yTweedie * gamma;
      end;
      output;
   end;
   run;
```

The following SAS statements invoke PROC GENMOD to fit the Tweedie GLM with the log link using all of the categorical and continuous variables. A Type III analysis is requested by the TYPE3 option in the MODEL statement.

```
   proc genmod data=tmp1;
      class C1-C5;
      model yTweedie = C1-C5 D1 D2 / dist=Tweedie type3;
   run;
```

The “Criteria For Assessing Goodness Of Fit” table is displayed in Output 51.12.1. The scaled Pearson $\chi^2$ is close to 1, indicating that the specified model fits the data well.

**Output 51.12.1** Tweedie Goodness of Fit Criteria

**The GENMOD Procedure**

**Criteria For Assessing Goodness Of Fit**

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| **Pearson Chi-Square** | 232 | 101.9124 | 0.4393 |
| **Scaled Pearson X2** | 232 | 251.5826 | 1.0844 |
| **Log Likelihood** | | -297.2106 | |
| **Full Log Likelihood** | | -297.2106 | |
| **AIC (smaller is better)** | | 634.4212 | |
| **AICC (smaller is better)** | | 638.0893 | |
| **BIC (smaller is better)** | | 704.8504 | |

The “LR Statistics For Type 3 Analysis” table is displayed in Output 51.12.2. As expected, the *p*-values for C3, C4, and d2 are not statistically significant at the 5% level.

**Output 51.12.2** Type III Analysis of Covariate Effects

**LR Statistics For Type 3 Analysis**

| Source | DF | Chi-Square | Pr > ChiSq |
|---|---|---|---|
| **c1** | 3 | 85.46 | <.0001 |
| **c2** | 3 | 48.18 | <.0001 |
| **c3** | 3 | 0.56 | 0.9050 |
| **c4** | 3 | 9.38 | 0.0247 |
| **c5** | 3 | 47.76 | <.0001 |
| **d1** | 1 | 0.00 | 0.9595 |
| **d2** | 1 | 1.31 | 0.2518 |

You can fix the power parameter for fitting the Tweedie GLM by using the P= option. The following SAS statements fit the model for C1, C2 and D1, while holding the power parameter at 1.5:

```
   proc genmod data=tmp1;
      class C1 C2;
      model yTweedie = C1 C2 D1 / dist=Tweedie(p=1.5) type3;
   run;
```

The parameter estimates are displayed in Output 51.12.3.

**Output 51.12.3** Tweedie Maximum Likelihood Parameter Estimates

**The GENMOD Procedure**

**Analysis Of Maximum Likelihood Parameter Estimates**

| Parameter | | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|
| **Intercept** | | 1 | 0.3440 | 0.1347 | 0.0801 | 0.6080 | 6.53 | 0.0106 |
| **c1** | 0 | 1 | -0.0722 | 0.1101 | -0.2880 | 0.1436 | 0.43 | 0.5120 |
| **c1** | 1 | 1 | -0.8952 | 0.1196 | -1.1296 | -0.6607 | 56.01 | <.0001 |
| **c1** | 2 | 1 | 0.0770 | 0.1073 | -0.1334 | 0.2873 | 0.51 | 0.4733 |
| **c1** | 3 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **c2** | 0 | 1 | 0.6138 | 0.1161 | 0.3862 | 0.8414 | 27.93 | <.0001 |
| **c2** | 1 | 1 | 0.5103 | 0.1150 | 0.2849 | 0.7356 | 19.70 | <.0001 |
| **c2** | 2 | 1 | 0.1001 | 0.1215 | -0.1380 | 0.3381 | 0.68 | 0.4099 |
| **c2** | 3 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **d1** | | 1 | -0.0211 | 0.1493 | -0.3136 | 0.2714 | 0.02 | 0.8876 |
| **Dispersion** | | 1 | 0.4951 | 0.0398 | 0.4172 | 0.5731 | | |
| **Power** | | 0 | 1.5000 | 0.0000 | 1.5000 | 1.5000 | | |

**Note:** The Tweedie dispersion parameter was estimated by maximum likelihood.
**Note:** The Tweedie power parameter was held fixed.

# References

Agresti, A. (2002). *Categorical Data Analysis*. 2nd ed. New York: John Wiley & Sons.

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

Akaike, H. (1979). "A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting." *Biometrika* 66:237–242.

Akaike, H. (1981). "Likelihood of a Model and Information Criteria." *Journal of Econometrics* 16:3–14.

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

Cameron, A. C., and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

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

Collett, D. (2003). *Modelling Binary Data*. 2nd ed. London: Chapman & Hall.

Cook, R. D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman & Hall.

Cox, D. R., and Snell, E. J. (1989). *The Analysis of Binary Data*. 2nd ed. London: Chapman & Hall.

Davison, A. C., and Snell, E. J. (1991). "Residuals and Diagnostics." In *Statistical Theory and Modelling*, edited by D. V. Hinkley, N. Reid, and E. J. Snell, 83–106. London: Chapman & 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 & Hall.

Dunn, P. K., and Smyth, G. K. (2005). "Series Evaluation of Tweedie Exponential Dispersion Model Densities." *Statistics and Computing* 15:267–280.

Dunn, P. K., and Smyth, G. K. (2008). "Series Evaluation of Tweedie Exponential Dispersion Model Densities by Fourier Inversion." *Statistics and Computing* 18:73–86.

Firth, D. (1991). "Generalized Linear Models." In *Statistical Theory and Modelling*, edited by D. V. Hinkley, N. Reid, and E. J. Snell, 55–82. London: Chapman & Hall.

Fischl, M. A., Richman, D. D., and Hansen, N. (1990). "The Safety and Efficacy of Zidovudine (AZT) in the Treatment of Subjects with Mildly Symptomatic Human Immunodeficiency Virus Type I (HIV) Infection." *Annals of Internal Medicine* 112:727–737.

Gamerman, D. (1997). "Sampling from the Posterior Distribution in Generalized Linear Models." *Statistics and Computing* 7:57–68.

Gilks, W. R. (2003). "Adaptive Metropolis Rejection Sampling (ARMS)." Software from MRC Biostatistics Unit, Cambridge, UK. `http://www.maths.leeds.ac.uk/~wally.gilks/adaptive.rejection/web_page/Welcome.html`.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). "Adaptive Rejection Metropolis Sampling within Gibbs Sampling." *Journal of the Royal Statistical Society, Series C* 44:455–472.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Gilks, W. R., and Wild, P. (1992). "Adaptive Rejection Sampling for Gibbs Sampling." *Journal of the Royal Statistical Society, Series C* 41:337–348.

Hardin, J. W., and Hilbe, J. M. (2003). *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC.

Hilbe, J. M. (1994). "Log Negative Binomial Regression Using the GENMOD Procedure in SAS/STAT Software." In *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, 1199–1204. Cary, NC: SAS Institute Inc. `https://support.sas.com/resources/papers/proceedings-archive/SUGI94/Sugi-94-205%20Hilbe.pdf`.

Hilbe, J. M. (2007). *Negative Binomial Regression*. New York: Cambridge University Press.

Hilbe, J. M. (2009). *Logistic Regression Models*. London: Chapman & Hall/CRC.

Hirji, K. F., Mehta, C. R., and Patel, N. R. (1987). "Computing Distributions for Exact Logistic Regression." *Journal of the American Statistical Association* 82:1110–1117.

Hougaard, P. (1986). "Survival Models for Heterogeneous Populations Derived from Stable Distributions." *Biometrika* 73:387–396.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999). "Monte Carlo EM for Missing Covariates in Parametric Regression Models." *Biometrics* 55:591–596.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag.

Ibrahim, J. G., and Laud, P. W. (1991). "On Bayesian Analysis of Generalized Linear Models Using Jeffreys' Prior." *Journal of the American Statistical Association* 86:981–986.

Lambert, D. (1992). "Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing." *Technometrics* 34:1–14.

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

Lawless, J. F. (2003). *Statistical Model and Methods for Lifetime Data*. 2nd ed. New York: John Wiley & Sons.

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

Lin, D. Y., Wei, L. J., and Ying, Z. (2002). "Model-Checking Techniques Based on Cumulative Residuals." *Biometrics* 58:1–12.

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

Littell, R. C., Freund, R. J., and Spector, P. C. (1991). *SAS System for Linear Models*. 3rd ed. Cary, NC: SAS Institute Inc.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

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

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

Meeker, W. Q., and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. New York: John Wiley & Sons.

Mehta, C. R., Patel, N. R., and Senchaudhuri, P. (1992). "Exact Stratified Linear Rank Tests for Ordered Categorical and Binary Data." *Journal of Computational and Graphical Statistics* 1:21–40.

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.

Muller, K. E., and Fetterman, B. A. (2002). *Regression and ANOVA: An Integrated Approach Using SAS Software*. Cary, NC: SAS Institute Inc.

Myers, R. H., Montgomery, D. C., and Vining, G. G. (2002). *Generalized Linear Models with Applications in Engineering and the Sciences*. New York: John Wiley & Sons.

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

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

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models*. 4th ed. Chicago: Irwin.

Pan, W. (2001). “Akaike’s Information Criterion in Generalized Estimating Equations.” *Biometrics* 57:120–125.

Pregibon, D. (1981). “Logistic Regression Diagnostics.” *Annals of Statistics* 9:705–724.

Preisser, J. S., and Qaqish, B. F. (1996). “Deletion Diagnostics for Generalised Estimating Equations.” *Biometrika* 83:551–562.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. 2nd ed. New York: John Wiley & Sons.

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.

Royall, R. M. (1986). “Model Robust Inference Using Maximum Likelihood Estimators.” *International Statistical Review* 54:221–226.

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

Simonoff, J. S. (2003). *Analyzing Categorical Data*. New York: Springer-Verlag.

Smyth, G. K. (1996). “Regression Analysis of Quantity Data with Exact Zeros.” In *Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering, Technology, and Management*, edited by R. J. Wilson, S. Osaki, and D. N. P. Murthy, 572–580. Queensland, Australia: Technology Management Centre, University of Queensland.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2002). “Bayesian Measures of Model Complexity and Fit.” *Journal of the Royal Statistical Society, Series B* 64:583–616. With discussion.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2000). *Categorical Data Analysis Using the SAS System*. 2nd ed. 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.

Tweedie, M. C. K. (1984). “An Index Which Distinguishes between Some Important Exponential Families.” In *Statistics: Applications and New Directions—Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, edited by J. K. Ghosh and J. Roy, 579–604. Calcutta: Indian Statistical Institute.

Ware, J. H., Dockery, S. 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.

White, H. (1982). “Maximum Likelihood Estimation of Misspecified Models.” *Econometrica* 50:1–25.

Williams, D. A. (1987). “Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions.” *Journal of the Royal Statistical Society, Series C* 36:181–191.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). “Models for Longitudinal Data: A Generalized Estimating Equation Approach.” *Biometrics* 44:1049–1060.