THE POWER TO KNOW.

SAS/STAT® 15.1 User's Guide

The GEE Procedure

This document is an individual chapter from *SAS/STAT® 15.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS/STAT® 15.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 15.1 User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2018

Chapter 47
The GEE Procedure

Contents

Overview: GEE Procedure

The GEE procedure implements the generalized estimating equations (GEE) approach (Liang and Zeger 1986), which extends the generalized linear model to handle longitudinal data (Stokes, Davis, and Koch 2012; Fitzmaurice, Laird, and Ware 2011; Diggle et al. 2002). For longitudinal studies, missing data are common, and they can be caused by dropouts or skipped visits. If missing responses depend on previous responses, the usual GEE approach can lead to biased estimates. So the GEE procedure also implements the weighted GEE method to handle missing responses that are caused by dropouts in longitudinal studies (Robins and Rotnitzky 1995; Preisser, Lohman, and Rathouz 2002). The GEE procedure in SAS/STAT 14.1 does not support the weighted GEE method for the multinomial distribution for polytomous responses.

The GEE method fits a marginal model to longitudinal data. The regression parameters in the marginal model are interpreted as population-averaged. For more information about the GEE method, see Fitzmaurice, Laird, and Ware (2011); Hardin and Hilbe (2003); Diggle et al. (2002); Lipsitz et al. (1994).

The GEE procedure compares most closely to the GENMOD procedure in SAS/STAT software. Both procedures implement the standard generalized estimating equation approach for longitudinal data; this approach is appropriate for complete data or when data are missing completely at random (MCAR). When the data are missing at random (MAR), the weighted GEE method produces valid inference. Molenberghs and Kenward (2007); Fitzmaurice, Laird, and Ware (2011); Mallinckrodt (2013); O'Kelly and Ratitch (2014) describe the weighted GEE method.

The GEE procedure includes alternating logistic regression (ALR) analysis for binary and ordinal multinomial responses. In ordinary GEEs, the association between pairs of responses are modeled with correlations. The ALR approach provides an alternative by using the log odds ratio to model the association between pairs. For more information about the log odds ratio and the ALR method, see the section "Alternating Logistic Regression" on page 3372. For binary responses the ALR algorithm of Carey, Zeger, and Diggle (1993) is implemented in both the GEE and GENMOD procedures. The GEE procedure also implements the ALR algorithm of Heagerty and Zeger (1996), which extends the ALR approach to ordinal multinomial responses. An ordinary GEE with the independent working correlation structure is also available for both nominal and ordinal multinomial data.

Getting Started

This section illustrates some of the basic features of the GEE procedure by analyzing longitudinal data from Stokes, Davis, and Koch (2012).

In this study, researchers followed 25 children at ages 8, 9, 10, and 11 years. The goal of this study is to investigate the health effects of air pollution on children. The binary response is the wheezing status of the children at four different ages. The explanatory variables are age, city, and passive smoking index (with values 0, 1, 2) that represented the degree of smoking in the home. The responses for individual children are assumed to be equally correlated, implying an exchangeable correlation structure.

The following statements create the data set Children:

```
data Children;
   input ID City$ @@;
   do i=1 to 4;
      input Age Smoke Symptom @@;
      output;
   end;
   datalines;
 1 steelcity  8 0 1  9 0 1  10 0 1  11 0 0
 2 steelcity  8 2 1  9 2 1  10 2 1  11 1 0
 3 steelcity  8 2 1  9 2 0  10 1 0  11 0 0
 4 greenhills 8 0 0  9 1 1  10 1 1  11 0 0
 5 steelcity  8 0 0  9 1 0  10 1 0  11 1 0
 6 greenhills 8 0 1  9 0 0  10 0 0  11 0 1
 7 steelcity  8 1 1  9 1 1  10 0 1  11 0 0
 8 greenhills 8 1 0  9 1 0  10 1 0  11 2 0
 9 greenhills 8 2 1  9 2 0  10 1 1  11 1 0
10 steelcity  8 0 0  9 0 0  10 0 0  11 1 0
11 steelcity  8 1 1  9 0 0  10 0 0  11 0 1
12 greenhills 8 0 0  9 0 0  10 0 0  11 0 0
13 steelcity  8 2 1  9 2 1  10 1 0  11 0 1
14 greenhills 8 0 1  9 0 1  10 0 0  11 0 0
15 steelcity  8 2 0  9 0 0  10 0 0  11 2 1
16 greenhills 8 1 0  9 1 0  10 0 0  11 1 0
17 greenhills 8 0 0  9 0 1  10 0 1  11 1 1
18 steelcity  8 1 1  9 2 1  10 0 0  11 1 0
19 steelcity  8 2 1  9 1 0  10 0 1  11 0 0
20 greenhills 8 0 0  9 0 1  10 0 1  11 0 0
21 steelcity  8 1 0  9 1 0  10 1 0  11 2 1
22 greenhills 8 0 1  9 0 1  10 0 0  11 0 0
23 steelcity  8 1 1  9 1 0  10 0 1  11 0 0
24 greenhills 8 1 0  9 1 1  10 1 1  11 2 1
25 greenhills 8 0 1  9 0 0  10 0 0  11 0 0
;
```

The following statements fit the model by the GEE method:

```
proc gee data=Children descending;
   class ID City;
   model Symptom = City Age Smoke / dist=bin link=logit;
   repeated subject=ID / type=exch covb corrw;
run;
```

Both the MODEL statement and the REPEATED statement are required.

The DIST=BIN and LINK=LOGIT options in the MODEL statement request a logistic regression with the variable Symptom as the response and City, Age, and Smoke as explanatory variables.

The REPEATED statement specifies the correlation structure and requests various tables in the output. The SUBJECT=ID option requests that individual subjects be identified in the input data set by the variable ID, which must be listed in the CLASS statement. Measurements of individual subjects at ages 8, 9, 10, and 11 are in the proper order in the data set, so the WITHIN= option is not required. The TYPE=EXCH option specifies an exchangeable working correlation structure, the COVB option requests the parameter estimate covariance matrix, and the CORRW option requests the working correlation matrix.

Figure 47.1 shows the “Model Information” table, which provides information about the specified logistic regression model and the input data set.

Figure 47.1 Model Information

The GEE Procedure

Model Information	
Data Set	WORK.CHILDREN
Distribution	Binomial
Link Function	Logit
Dependent Variable	Symptom

Figure 47.2 displays general information about the GEE analysis. Each subject has four measurements.

Figure 47.2 GEE Model Information

GEE Model Information	
Correlation Structure	Exchangeable
Subject Effect	ID (25 levels)
Number of Clusters	25
Correlation Matrix Dimension	4
Maximum Cluster Size	4
Minimum Cluster Size	4

Figure 47.3 displays the model-based and empirical covariance matrices of the parameter estimates.

Figure 47.3 Covariance Matrices of Parameter Estimates

Covariance Matrix (Model-Based)

	Prm1	Prm2	Prm4	Prm5
Prm1	3.26069	-0.16313	-0.32274	-0.12257
Prm2	-0.16313	0.24015	0.002520	0.03422
Prm4	-0.32274	0.002520	0.03379	0.004471
Prm5	-0.12257	0.03422	0.004471	0.09533

Covariance Matrix (Empirical)

	Prm1	Prm2	Prm4	Prm5
Prm1	4.09770	-0.55261	-0.37280	-0.29397
Prm2	-0.55261	0.29538	0.03719	0.09143
Prm4	-0.37280	0.03719	0.03550	0.02064
Prm5	-0.29397	0.09143	0.02064	0.07957

The exchangeable working correlation matrix is displayed in Figure 47.4.

Figure 47.4 Working Correlation Matrix

Working Correlation Matrix				
	Obs 1	**Obs 2**	**Obs 3**	**Obs 4**
Obs 1	1.0000	0.0883	0.0883	0.0883
Obs 2	0.0883	1.0000	0.0883	0.0883
Obs 3	0.0883	0.0883	1.0000	0.0883
Obs 4	0.0883	0.0883	0.0883	1.0000

The parameter estimates table, shown in Figure 47.5, 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. You can create a table that uses model-based standard errors by specifying the MODELSE option in the REPEATED statement. The results indicate that smoking exposure is significant with a p-value of 0.0211, Age is marginally influential with a p-value of 0.0893, and City does not influence wheezing. The parameter estimate for Age is –0.3201, which indicates that the odds ratio of wheezing for the children at the higher age group compared to those in the lower age group is $e^{-0.3201} = 0.726$.

Figure 47.5 GEE Parameter Estimates Table

Parameter Estimates for Response Model with Empirical Standard Error Estimates							
Parameter		**Estimate**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
Intercept		2.2615	2.0243	-1.7060	6.2290	1.12	0.2639
City	greenhil	0.0418	0.5435	-1.0234	1.1070	0.08	0.9387
City	steelcit	0.0000	0.0000	0.0000	0.0000	.	.
Age		-0.3201	0.1884	-0.6894	0.0492	-1.70	0.0893
Smoke		0.6506	0.2821	0.0978	1.2035	2.31	0.0211

Goodness-of-fit criteria for the model are displayed in Figure 47.6. For more information about the quasi-likelihood information criterion (QIC), see the section "Quasi-likelihood Information Criterion" on page 3371.

Figure 47.6 Model Fit Criteria

GEE Fit Criteria	
QIC	137.1373
QICu	136.2173

Syntax: GEE Procedure

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

PROC GEE < *options* > ;
  BY *variables* ;
  CLASS *variable* < **(***options***)** > . . . < *variable* < **(***options***)** > > < / *options* > ;
  EFFECTPLOT < *plot-type* < **(***plot-definition-options***)** > > < / *options* > ;
  ESTIMATE < '*label*' > *estimate-specification* < / *options* > ;
  FREQ | FREQUENCY *variable* ;
  LSMEANS < *model-effects* > < / *options* > ;
  LSMESTIMATE *model-effect* < '*label*' > *values* < *divisor*=*n* > < **,** . . . < '*label*' > *values* < *divisor*=*n* > >
      < / *options* > ;
  MISSMODEL < *effects* > < / *options* > ;
  MODEL *response* **=** < *effects* > < / *options* > ;
  OUTPUT < **OUT=***SAS-data-set* > < *keyword*=*name* . . . *keyword*=*name* > ;
  REPEATED SUBJECT=*subject-effect* < / *options* > ;
  SLICE *model-effect* < / *options* > ;
  STORE < **OUT=** >*item-store-name* < / **LABEL=**'*label*' > ;
  WEIGHT *variable* ;

The syntax of the GEE procedure compares most closely to that of the GENMOD procedures. The PROC GEE, MODEL, and REPEATED statements are required. All other statements can appear only once. The following sections describe the PROC GEE statement and then describe the other statements in alphabetical order.

PROC GEE Statement

PROC GEE < *options* > ;

The PROC GEE statement invokes the GEE procedure. Table 47.1 summarizes the *options* available in the PROC GEE statement.

Table 47.1 PROC GEE Statement Options

Option	Description
DATA=	Specifies the input data set
DESCENDING	Sorts the response variable in the reverse of the default order
NAMELEN=	Specifies the length of effect names
ORDER=	Specifies the sort order of CLASS variable
PLOTS	Controls the plots that are produced through ODS Graphics

You can specify the following *options*.

DATA=*SAS-data-set*

specifies the SAS data set that contains the data to be analyzed. If you omit the DATA= option, PROC GEE uses the most recently created SAS data set.

DESCENDING

DESCEND

DESC

requests that the levels of the response variable for the binomial model that uses a single-variable response syntax be sorted in the reverse of the default order.

NAMELEN=*number*

specifies the length to which long effect names are shortened. The default and minimum value is 20.

PLOTS < = *plot-request* >

controls the plots produced through ODS Graphics. For example:

```
proc gee plots=histogram;
   model y=x1;
run;
```

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 623 in Chapter 21, "Statistical Graphics Using ODS."

You can specify the following *plot-requests*:

ALL

requests that all default plots be produced.

HISTOGRAM

creates a histogram for the predicted weights from the missingness model.

NONE

suppresses all plots.

BY Statement

BY *variables* **;**

You can specify a BY statement in PROC GEE 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 GEE 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 discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

CLASS Statement

CLASS *variables < / options >* **;**

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

Classification variables can be either character or numeric. CLASS levels are determined from the formatted values of the *variables*. Thus, you can use formats to group values into levels. For more information, see the discussion of the FORMAT procedure in the *Base SAS Procedures Guide* and the discussions of the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*.

You can specify the following *options* for classification *variables*:

DESCENDING

DESC

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

ORDER=*order-type*

specifies the sort order for the categories of categorical variables. This ordering determines which parameters in the model correspond to each level in the data. When the default ORDER=FORMATTED is in effect for numeric variables for which you have supplied no explicit format, the levels are ordered by their internal values. Table 47.2 shows how PROC GEE interprets values of the ORDER= option.

Table 47.2 Sort Order for Categorical Variables

order-type	**Levels Sorted By**
DATA	Order of appearance in the input data set
FORMATTED	External formatted value, except for numeric variables that have no explicit format, which are sorted by their unformatted (internal) value
FREQ	Descending frequency count; levels that have the most observations come first in the order
FREQDATA	Order of descending frequency count, and within counts by order of appearance in the input data set when counts are tied
FREQFORMATTED	Order of descending frequency count, and within counts by formatted value (as above) when counts are tied
FREQINTERNAL	Order of descending frequency count, and within counts by unformatted value when counts are tied
INTERNAL	Unformatted value

For the FORMATTED and INTERNAL values, the sort order is machine-dependent. If you specify the ORDER= option in the MODEL statement and the ORDER= option in the CLASS statement, the former takes precedence.

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 Language Reference: Concepts*.

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 47.3 describes the available *plot-types* and their *plot-definition-options*.

Table 47.3 *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
	X= CLASS variable or effect

CONTOUR
Displays a contour plot of predicted values against two continuous covariates.
	PLOTBY= variable or CLASS effect
	X= continuous variable
	Y= continuous variable

FIT
Displays a curve of predicted values versus a continuous variable.
	PLOTBY= variable or CLASS effect
	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
	SLICEBY= variable or CLASS effect
	X= CLASS variable or effect

MOSAIC
Displays a mosaic plot of predicted values using up to three CLASS effects.
	PLOTBY= variable or CLASS effect
	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
	SLICEBY= variable or CLASS effect
	X= continuous variable

For full details about the syntax and options of the EFFECTPLOT statement, see the section "EFFECTPLOT Statement" on page 423 in Chapter 19, "Shared Concepts and Topics."

ESTIMATE Statement

ESTIMATE < '*label*' > *estimate-specification* < **(***divisor***=***n***)** >
< **,** . . . < '*label*' > *estimate-specification* < **(***divisor***=***n***)** > >
< / *options* > **;**

The ESTIMATE statement provides a mechanism for obtaining custom hypothesis tests. Estimates are formed as linear estimable functions of the form $\mathbf{L}\boldsymbol{\beta}$. You can perform hypothesis tests for the estimable functions, construct confidence limits, and obtain specific nonlinear transformations.

Table 47.4 summarizes the *options* available in the ESTIMATE statement.

Table 47.4 ESTIMATE Statement Options

Option	Description
Construction and Computation of Estimable Functions	
DIVISOR=	Specifies a list of values to divide the coefficients
NOFILL	Suppresses the automatic fill-in of coefficients for higher-order effects
SINGULAR=	Tunes the estimability checking difference
Degrees of Freedom and *p*-Values	
ADJUST=	Determines the method of multiple comparison adjustment of estimates
ALPHA=α	Determines the confidence level $(1-\alpha)$
LOWER	Performs one-sided, lower-tailed inference
STEPDOWN	Adjusts multiplicity-corrected *p*-values further in a step-down fashion
TESTVALUE=	Specifies values under the null hypothesis for tests
UPPER	Performs one-sided, upper-tailed inference
Statistical Output	
CL	Constructs confidence limits
CORR	Displays the correlation matrix of estimates
COV	Displays the covariance matrix of estimates
E	Prints the **L** matrix
JOINT	Produces a joint *F* or chi-square test for the estimable functions
PLOTS=	Produces ODS statistical graphics if the analysis is sampling-based
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 estimates
ILINK	Computes and displays estimates and standard errors on the inverse linked scale

For more information about the syntax of the ESTIMATE statement, see the section "ESTIMATE Statement" on page 451 in Chapter 19, "Shared Concepts and Topics."

FREQ Statement

FREQ *variable* **;**

FREQUENCY *variable* **;**

The *variable* in the FREQ statement identifies a variable in the input data set that contains the frequency of occurrence of each observation. PROC GEE treats each observation as if it appeared *n* times, where *n* is the value of the FREQ variable for the observation. If the frequency value is not an integer, it is truncated to an integer. If it is less than 1 or missing, the observation is not used. The frequencies must be the same for all observations within each subject.

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 47.5 summarizes the *options* available in the LSMEANS statement.

Table 47.5 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=α	Determines the confidence level $(1 - \alpha)$
STEPDOWN	Adjusts multiple-comparison *p*-values further in a step-down fashion
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

Table 47.5 *continued*

Option	Description
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
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 more information about the syntax of the LSMEANS statement, see the section "LSMEANS Statement" on page 467 in Chapter 19, "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 47.6 summarizes the *options* available in the LSMESTIMATE statement.

Table 47.6 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

Table 47.6 *continued*

Option	Description
Degrees of Freedom and *p*-Values	
ADJUST=	Determines the method of multiple-comparison adjustment of LS-means differences
ALPHA=α	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
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 487 in Chapter 19, "Shared Concepts and Topics."

MISSMODEL Statement

MISSMODEL *effects < / options >* **;**

The MISSMODEL statement requests a weighted GEE analysis. It specifies a logistic regression that is used to estimate the weights under the MAR assumption. If the pattern of missing data is intermittent (not dropout), the GEE procedure terminates and does not perform an analysis.

You can use the same effects or different effects in the MODEL and MISSMODEL statements. Explanatory variables can be continuous or classification variables. Classification variables can be character or numeric. Explanatory variables that represent nominal (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 effects is the same as for the GLM procedure. For more information, see the section "Specification of Effects" on page 4020 in Chapter 50, "The GLM Procedure."

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

MAXWEIGHT=*number*

truncates the predicted weights from the missingness model if they are larger than *number*, where *number* ≥ 1.

TYPE=OBSLEVEL | SUBLEVEL

specifies the type of weighted GEE method. You can specify the following values:

OBSLEVEL specifies the observation-level weighted GEE method.

SUBLEVEL specifies the subject-level weighted GEE method.

By default, TYPE=OBSLEVEL.

MODEL Statement

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

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

The MODEL statement specifies the response (dependent variable) and the effects (explanatory variables). If you omit the explanatory variables, PROC GEE fits an intercept-only model. An intercept term is included in the model by default. You can remove the intercept by specifying the NOINT option.

You can specify the response in the form of a single variable (*response*) or in the form of a ratio of two variables (*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: one for the event counts and one for trial counts. These two variables are separated by a slash (/). The value of the *events* variable must be nonnegative, and the value of the *trials* variable must be equal to or greater than the value of the *events* variable for an observation to be valid. The *events* and *trials* variables can take non-integer values.

When each observation in the input data set contains a single trial from a binomial experiment, use the *response* form of the MODEL statement. The response variable can be numeric or character. The ordering of response levels is critical in these models.

Responses for the Poisson distribution must be all nonnegative, but they can be non-integer 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 that represent nominal (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 specifying effects is the same as for the GLM procedure. For more information, see the section "Specification of Effects" on page 4020 in Chapter 50, "The GLM Procedure."

Table 47.7 summarizes the *options* available in the MODEL statement.

Table 47.7 MODEL Statement Options

Option	Description
ALPHA=	Sets the confidence coefficient
DIST=	Specifies the probability distribution
LINK=	Specifies the link function
NOINT	Requests no intercept term
NOSCALE	Holds the scale parameter fixed
OFFSET=	Specifies a variable in the input data set to be used as an offset
SCALE=	Specifies the value used for the scale
TYPE3	Computes statistics for Type 3 contrasts
WALD	Requests Wald statistics for Type 3 contrasts

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

ALPHA=*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.

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 the LINK= option, a default link function is chosen as displayed in Table 47.8. If you specify neither the DIST= option nor the LINK= option, then the GEE procedure defaults to the normal distribution with the identity link function.

Table 47.8 Distributions and Default Link Functions

DIST=	Distribution	Default Link Function
BINOMIAL | BIN | B	Binomial	Logit
GAMMA | GAM | G	Gamma	Reciprocal
IGAUSSIAN | IG	Inverse Gaussian	Reciprocal square
MULTINOMIAL | MULT	Multinomial	Cumulative logit
NEGBIN | NB	Negative binomial	Log
NORMAL | NOR | N	Normal	Identity
POISSON | POI | P	Poisson	Log

LINK=*keyword*

specifies the link function in the model. You can specify the *keywords* shown in Table 47.9.

Table 47.9 Built-In Link Functions of the GEE Procedure

LINK=	Link Function	$g(\mu) = \eta =$
CLOGLOG | CLL	Complementary log-log	$\log(-\log(1-\mu))$
CUMCLL | CCLL	Cumulative complementary log-log	$\log(-\log(1-\pi))$
CUMLOGIT| CLOGIT	Cumulative logit	$\log(\pi/(1-\pi))$
CUMPROBIT | CPROBIT	Cumulative probit	$\Phi^{-1}(\pi)$
GLOGIT	Generalized logit	
IDENTITY | ID	Identity	μ
LOG	Log	$\log(\mu)$
LOGIT	Logit	$\log(\mu/(1-\mu))$
PROBIT	Probit	$\Phi^{-1}(\mu)$
INVERSE | RECIPROCAL	Reciprocal	$1/\mu$
POWERMINUS2	Power with exponent –2	$1/\mu^2$

For the probit and cumulative probit links, $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution. If you do not specify the LINK= option, then by default the 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 with ordinal responses, with cumulative probabilities indicated by π. The GLOGIT link function is appropriate only for the multinomial distribution with nominal responses.

NOINT

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

NOSCALE

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

OFFSET=*variable*

specifies a variable in the input data set to be used as an offset variable. This variable cannot be a CLASS variable, the response variable, or any of the explanatory variables.

SCALE=*number*

SCALE=PEARSON | P

PSCALE

SCALE=DEVIANCE | D

DSCALE

specifies the value used for the scale parameter when 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. 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 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.

TYPE3

requests that statistics for Type 3 contrasts be computed for each effect specified in the MODEL statement. The default analysis is to compute score statistics for the contrasts. Type 3 analyses using the score statistics are not supported for nominal response data or weighted GEE methods. 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.

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, predicted values of the mean, and confidence limits for predicted values.

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.

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).

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

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(Y $\leq$ _LEVEL_), where Y is the response variable).

RESCHI represents the Pearson (chi) residual for identifying observations that are poorly accounted for by the model. This option is not available for the multinomial distribution.

RESRAW represents the raw residual for identifying poorly fitted observations. This option is not available for the multinomial distribution.

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

REPEATED Statement

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

The REPEATED statement specifies the correlation structure of the responses for GEE model fitting. In addition, the REPEATED statement controls the iterative fitting algorithm and specifies optional output.

Table 47.10 summarizes the *options* available in the REPEATED statement.

Table 47.10 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

Table 47.10 *continued*

Option	Description
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 an initial value of the intercept
LOGOR=	Specifies the use of alternating logistic regression and a model for the log odds ratio
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 a parameter estimates table with the model-based standard errors
SUBCLUSTER=	Specifies a variable that defines subclusters
SUBJECT=	Identifies a different subject (cluster)
TYPE=	Specifies the working correlation matrix structure
WITHIN=	Specifies the order of measurements within subjects
ZDATA=	Specifies the full **z** matrix
ZROW=	Specifies the rows of the **z** matrix

You must specify the SUBJECT= option:

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 (level) of the effect identifies a different subject (cluster). Responses from different subjects are assumed to be statistically independent, and responses within subjects are assumed to be correlated. You must specify a *subject-effect*, and you must list variables that are used in defining the *subject-effect* in the CLASS statement.

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

ALPHAINIT=*numbers*

specifies initial values for log odds ratio regression parameters if you specify the option LOGOR= for data that have either binary or ordinal multinomial responses. The default value of *numbers* is 0.01.

CONVERGE=*number*

specifies the convergence criterion for GEE parameter estimation. If the maximum absolute difference between regression parameter estimates is less than *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.

CORRB

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

CORRW

displays the estimated working correlation matrix. If you specify TYPE=EXCH for the exchangeable working correlation structure, then the CORRW option is not needed to view the estimated correlation, because a table that contains the single estimated correlation is printed by default.

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.

INITIAL=*numbers*

specifies initial values of the regression parameters estimation, other than the intercept parameter, for GEE estimation. If you do not specify this option, then the estimated regression parameters (assuming independence for all responses) are used for the initial values.

INTERCEPT=*number*

specifies an initial value of the intercept regression parameter in the GEE model.

LOGOR=*log-odds-ratio-structure-keyword*

specifies the use of the alternating logistic regression (ALR) method and the regression model structure for the log odds ratio. For data that have either a binary or ordinal multinomial response distribution, the ALR method uses the log odds ratio to model the association of the responses from subjects. For more information about the ALR method and examples of specifying log odds ratio models, see the section "Alternating Logistic Regression" on page 3372. You can specify the values that are shown in Table 47.11.

Table 47.11 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 **z** matrix specified in ZDATA= data set
ZREP	Single cluster specification for replicated **z** matrix specified in ZDATA= data set
ZREP(*matrix*)	Single cluster specification for replicated **z** matrix

For ordinal multinomial data, only the exchangeable regression structure that is specified by LOGOR=EXCH is supported. You should specify the option LOGOR= or TYPE=, but not both.

MAXITER=*number*

MAXIT=*number*

specifies the maximum *number* of iterations allowed in the iterative GEE estimation process. By default, MAXITER=50.

MCORRB

displays the estimated regression parameter model-based correlation matrix.

MCOVB

displays the estimated regression parameter model-based covariance matrix.

MODELSE

displays a parameter estimates table that uses model-based standard errors for inference. By default, a "Parameter Estimates" table that is based on empirical standard errors is displayed.

SUBCLUSTER=*variable*

SUBCLUST=*variable*

specifies a *variable* that defines subclusters for the 1-nested or k-nested log odds ratio association modeling structures for data that have a binary response distribution. A 1-nested or k-nested modeling structure is specified in the option LOGOR=, and *variable* must be listed in the CLASS statement. For definitions of the 1-nested and k-nested modeling structures, see the section "Specifying Log Odds Ratio Models" on page 3374.

TYPE=*correlation-structure-keyword*

CORR=*correlation-structure-keyword*

specifies the structure of the working correlation matrix that is used to model the correlation of the responses from subjects for ordinary GEEs. You can specify the values that are shown in Table 47.12 (for definitions of the correlation matrix types, see Table 47.13 in the section "Details: GEE Procedure" on page 3369).

Table 47.12 Correlation Structure Types

Keyword	**Correlation Structure Type**
AR | **AR**(1)	Autoregressive(1)
EXCH | **CS**	Exchangeable
IND	Independent
MDEP(*number*)	m-dependent, where m = *number*
UNSTR | **UN**	Unstructured
USER(*matrix*) | **FIXED**(*matrix*)	Fixed, user-specified correlation matrix

For example, the following option specifies a fixed 4 × 4 correlation matrix:

```
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 )
```

By default, TYPE=IND. When you specify the alternating logistic regression method using the option LOGOR= you should not specify TYPE=.

WITHINSUBJECT=*within-subject-effect*

WITHIN=*within-subject-effect*

defines an effect that specifies 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 you do not specify the WITHIN= option for the standard GEE method, missing values are assumed to be the last values and are not used; the remaining observations are then ordered in the sequence in which they are provided in the input data set. If you do not specify the WITHIN= option for the weighted GEE method, the observations are assumed to be ordered in the sequence in which they are provided in the input data set.

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

ZDATA=*SAS-data-set*

specifies a SAS data set that contains either the full **z** matrix for log odds ratio association modeling for data with binary responses 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 that correspond to rows of the **z** matrix for log odds ratio association modeling for data with binary responses.

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 19.23 in Chapter 19, “Shared Concepts and Topics.” For more information about the syntax of the SLICE statement, see the section “SLICE Statement” on page 516 in Chapter 19, “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 520 in Chapter 19, “Shared Concepts and Topics.”

WEIGHT Statement

WEIGHT *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.

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

Details: GEE Procedure

Generalized Estimating Equations

The marginal model is commonly used in analyzing longitudinal data when the population-averaged effect is of interest. To estimate the regression parameters in the marginal model, Liang and Zeger (1986) proposed the generalized estimating equations method, which is widely used.

Suppose $y_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, K$, represent the *j*th response of the *i*th subject, which has a vector of covariates x_{ij}. There are n_i measurements on subject *i*, and the maximum number of measurements per subject is *T*.

Suppose the responses of the *i*th subject be $\mathbf{Y}_i = [y_{i1}, \ldots, y_{in_i}]'$ with corresponding means $\boldsymbol{\mu}_i = [\mu_{i1}, \ldots, \mu_{in_i}]'$. For generalized linear models, the marginal mean μ_{ij} of the response y_{ij} is related to a linear predictor through a link function $g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}$, and the variance of y_{ij} depends on the mean through a variance function $v(\mu_{ij})$.

An estimate of the parameter $\boldsymbol{\beta}$ in the marginal model can be obtained by solving the generalized estimating equations,

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

where $\mathbf{V}_i$ is the working covariance matrix of $\mathbf{Y}_i$.

Only the mean and the covariance of $\mathbf{Y}_i$ are required in the GEE method; a full specification of the joint distribution of the correlated responses is not needed. This is particularly convenient because the joint distribution for noncontinuous responses involves high-order associations and is complicated to specify. Moreover, the regression parameter estimates are consistent even when the working covariance is incorrectly specified. Because of these properties, the GEE method is popular in situations where the marginal effect is of interest and the responses are not continuous. However, the GEE approach can lead to biased estimates when missing responses depend on previous responses. The weighted GEE method, which is described in the section "Weighted Generalized Estimating Equations under the MAR Assumption" on page 3376, can provide unbiased estimates.

Working Correlation Matrix

Suppose $\mathbf{R}_i(\boldsymbol{\alpha})$ is 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 whose *j*th diagonal element is $v(\mu_{ij})$ and $\mathbf{W}_i$ is an $n_i \times n_i$ diagonal matrix whose *j*th diagonal is w_{ij}, where w_{ij} is a weight variable that is specified in 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 $\mathbf{Y}_i$.

In practice, 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 matrix as $\mathbf{R}_0 = \mathbf{I}$, which is the identity matrix, the GEE reduces to the independence estimating equation.

Table 47.13 shows the working correlation structures that are supported by the GEE procedure and the estimators that are used to estimate the working correlations.

Table 47.13 Working Correlation Structures and Estimators

Working Correlation Structure	Estimator
Fixed $\text{Corr}(Y_{ij}, Y_{ik}) = r_{jk}$ 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** $\text{Corr}(Y_{ij}, Y_{i,j+t}) = \begin{cases} 1 & t = 0 \\ \alpha_t & t = 1, 2, \ldots, m \\ 0 & t > m \end{cases}$	$\hat{\alpha}_t = \frac{1}{(K_t - p)\phi} \sum_{i=1}^{K} \sum_{j \le n_i - t} e_{ij} e_{i,j+t}$ $K_t = \sum_{i=1}^{K} (n_i - t)$
Exchangeable $\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha & j \neq k \end{cases}$	$\hat{\alpha} = \frac{1}{(N^* - p)\phi} \sum_{i=1}^{K} \sum_{j<k} e_{ij} e_{ik}$ $N^* = 0.5 \sum_{i=1}^{K} n_i (n_i - 1)$
Unstructured $\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha_{jk} & j \neq k \end{cases}$	$\hat{\alpha}_{jk} = \frac{1}{(K - p)\phi} \sum_{i=1}^{K} e_{ij} e_{ik}$

Table 47.13 *continued*

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

Dispersion Parameter

The dispersion parameter ϕ is estimated by

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

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

The square root of $\hat{\phi}$ is reported by PROC GEE as the scale parameter in the "Parameter Estimates for Response Model with Model-Based Standard Error" output table. If a fixed scale parameter is specified by using the NOSCALE option in the MODEL statement, then the fixed value is used in estimating the model-based covariance matrix and standard errors.

Quasi-likelihood Information Criterion

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

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

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

where the quasi-likelihood contribution of the jth observation in the ith cluster is defined in the section "Quasi-likelihood Functions" on page 3372 and $\hat{\boldsymbol{\beta}}(R)$ are the parameter estimates that are obtained by using the GEE approach with the working correlation of interest R.

QIC is defined as

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

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 $\hat{\boldsymbol{\beta}}(R)$, which are the parameter estimates that are obtained by using the GEE approach with the working correlation of interest R.

PROC GEE also computes an approximation to $\text{QIC}(R)$, which is defined by Pan (2001) as

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

where p is the number of regression parameters.

Pan (2001) notes that QIC is appropriate for selecting regression models and working correlations, whereas 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 that is evaluated at the regression parameters $\boldsymbol{\beta}$ is expressed by $Q(\boldsymbol{\beta}, \phi; (Y_{ij}, \mathbf{X}_{ij})) = \frac{Q_{ij}}{\phi}$, where Q_{ij} is defined in the following list. These definitions are used in the computation of the quasi-likelihood information criteria (QIC) for goodness of fit of models that are fit by the GEE approach. The w_{ij} are prior weights, if any, that are specified in the WEIGHT or FREQ statement. 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):

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

Alternating Logistic Regression

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 by using log odds ratios instead of using correlations, as ordinary GEEs do. The ALR algorithm of Heagerty and Zeger (1996) extends the method to GEEs that have ordinal multinomial responses (that is, they fall into one of C ordered categories).

ALR for Binary Data

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, because $\mu_{ij} = \Pr(Y_{ij} = 1)$:

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

Therefore, the correlation 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 γ_{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 by specifying 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.

ALR for Ordinal Multinomial Data

For ordinal multinomial data, let $O_{ij}, i = 1, \ldots, K, j = 1, \ldots, n_i$, denote the *j*th measurement on the *i*th subject. To apply the ALR algorithm, the responses O_{ij} are represented by a vector $\mathbf{Y}_{ij} = \left[Y_{ij1}, \ldots, Y_{ijC-1}\right]'$ of cumulative indicator variables $Y_{ijc} = \mathrm{I}(O_{i,j} \le c)$. You model the cumulative probabilities $\mu_{ijc} = E\left(Y_{ijc}\right)$ by using a cumulative link function,

$$g\left(\mu_{ijc}\right) = \boldsymbol{\beta}_c + \mathrm{x}_{ij}'\boldsymbol{\beta}, \text{ for } c = 1, \ldots, C-1$$

where $\beta_1, \beta_2, \ldots, \beta_{C-1}$ are increasing intercept terms that depend only on the level c. Let the binary vector that represents the responses of the *i*th subject be $\mathbf{Y}_i = \left[\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{in_i}\right]'$ with corresponding means $\boldsymbol{\mu}_i = \left[\mu_{i1}, \ldots, \mu_{in_i}\right]'$.

The log odds ratio between two indicator variables Y_{ijc_1} and Y_{ikc_2} is modeled as

$$\gamma_{i(jk)(c_1c_2)} = \log(\mathrm{OR}(Y_{ijc_1}, Y_{ikc_2})) = \mathbf{z}_{i(jk)(c_1c_2)}'\boldsymbol{\alpha}$$

for $q \times 1$ regression parameters $\boldsymbol{\alpha}$ and fixed coefficients $\mathbf{z}_{i(jk)(c_1c_2)}$. As in Carey, Zeger, and Diggle (1993), $\boldsymbol{\alpha}$ then provides a vector of regression parameters in a logistic model for the conditional expectation $\xi_{i(jk)(c_1c_2)} = E\left(Y_{ijc_1}|Y_{ikc_2}\right)$. To estimate $\boldsymbol{\alpha}$, the conditional expectation is considered for all pairs Y_{ijc_1} and Y_{ikc_2} with $j < k$. Let

$$\boldsymbol{\xi}_{i(jk)} = \left[\xi_{i(jk)(11)}, \xi_{i(jk)(12)}, \ldots, \xi_{i(jk)(21)}, \ldots, \xi_{i(jk)(C-1,C-1)}\right]'$$

$$\boldsymbol{\xi}_{i} = \left[\boldsymbol{\xi}_{i(12)}, \boldsymbol{\xi}_{i(13)}, \ldots, \boldsymbol{\xi}_{i(23)}, \ldots, \boldsymbol{\xi}_{i(n_i-1n_i)}\right]'$$

$$\mathbf{Y}_i^* = \Big[\overbrace{Y_{i1} \otimes e_{C-1}, \ldots, Y_{i1} \otimes e_{C-1}}^{n_i-1}, \underbrace{Y_{i2} \otimes e_{C-1}, \ldots, Y_{i2} \otimes e_{C-1}}_{n_i-2}, \ldots, \overbrace{Y_{in_i-1} \otimes e_{C-1}}^{1}\Big]'$$

where $\otimes$ denotes the Kronecker product and e_l denotes a vector of dimension l composed of ones. The difference $\mathbf{Y}_i^* - \boldsymbol{\xi}_i$ represents the residuals of the model for the conditional expectation.

For both binary and multinomial data, the ALR estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the simultaneous solutions to the estimating equations

$$\mathrm{S}_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \mathbf{V}_{i11}^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) = \mathbf{0}$$

$$\mathrm{S}_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\xi}_i}{\partial \boldsymbol{\alpha}}' \mathbf{V}_{i33}^{-1} \left(\mathbf{Y}_i^* - \boldsymbol{\xi}_i\right) = \mathbf{0}$$

where $\mathbf{V}_{i11} = \mathrm{cov}\left(\mathbf{Y}_i\right)$ and $\mathbf{V}_{i33} = \mathrm{diag}\left[\boldsymbol{\xi}_i\left(1 - \boldsymbol{\xi}_i\right)\right]$. The fitting 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 the rows of the matrix $\mathbf{z}$. For binary data, there is a row $\mathbf{z}_{ijk}$ for each cluster i and within-cluster pair (j, k). For ordinal multinomial data, there is a row $\mathbf{z}_{i(jk)(c_1c_2)}$ for each cluster i, within-cluster pair (j, k), and choice of levels (c_1, c_2).

For ordinal multinomial data, the GEE procedure supports only the ALR method that uses a fully exchangeable regression structure for the log odds ratio. In a fully exchangeable model, the log odds ratio is constant for all clusters i, within-cluster pair (j, k), and levels (c_1, c_2). You select a fully exchangeable model for the log odds ratio by specifying LOGOR=EXCH.

For binary data, the GEE procedure provides several methods of specifying $\mathbf{z}_{ijk}$. You apply these methods by specifying 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 α 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 that takes a different value for k treatment centers, then when you specify LOGOR=LOGORVAR(Center), you get a model that has 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 GEE computes a log odds ratio parameter for pairs that have 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 that have the same value of *variable* correspond to one parameter; pairs that have different values of *variable* correspond to the other parameter. For example, if patients are clustered by hospital and subclusters are the wards within those hospitals, then the outcomes of patients within the same ward have one log odds ratio parameter, and the outcomes of patients from different wards have the other parameter.

ZFULL specifies the full $\mathbf{z}$ matrix. You must also specify a SAS data set that contains the $\mathbf{z}$ matrix by using 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 **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 that you specify in the REPEATED statement for the SUBJECT effect must also be present in the ZDATA= data set to identify clusters. You must specify variables in the data set that define the columns of the **z** matrix by using 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 that indicate the cluster pairs corresponding to each row of the **z** matrix by using 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 47.4 for an example of specifying a full **z** matrix.

ZREP

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

ZREP(*matrix*)

specifies direct input of the replicated **z** matrix. You specify the **z** matrix for one cluster by using the syntax LOGOR=ZREP ($(y_j \;\; y_k)\; z_{jk1} \;\; z_{jk2} \cdots z_{jkq}, \cdots$), where y_j and y_k are numbers that represent a pair of observations from the ith cluster and the values $z_{jk1}, z_{jk2}, \ldots, z_{jkq}$ make up the corresponding row $\mathbf{z}_{ijk}$ of the **z** matrix. The number of specified rows is $\frac{n_{max}(n_{max}-1)}{2}$, where n_{max} is the size of a complete cluster (a cluster with no missing observations). For example,

```
logor =  zrep((1 2) 1 0,
              (1 3) 1 0,
              (1 4) 1 0,
              (2 3) 1 1,
              (2 4) 1 1,
              (3 4) 1 1)
```

specifies the $\frac{4\times 3}{2} = 6$ rows of the **z** matrix for a cluster of size 4 with $q = 2$ log odds ratio parameters. The log odds ratio for the pairs (1 2), (1 3), (1 4) is α_1, and the log odds ratio for the pairs (2 3), (2 4), (3 4) is $\alpha_1 + \alpha_2$.

Weighted Generalized Estimating Equations under the MAR Assumption

In longitudinal studies, response measurements are often missing because of skipped visits or dropouts. Suppose r_{ij} is the indicator that the response y_{ij} is observed, where $r_{ij} = 1$ if y_{ij} is observed and 0 otherwise. Missing data patterns can be classified into two types: dropout and intermittent. A dropout occurs if an individual skips a particular visit and then never comes back for subsequent visits. That is, if $r_{ij} = 0$, then $r_{ik} = 0$ for all $k > j$. Otherwise, the missing data pattern is intermittent. Intermittent patterns can be quite complicated; only dropout patterns are considered here.

The mechanism for missingness can be described by a statistical model for the probability of observing a missing value, and making the right assumption about the mechanism is crucial to methods that handle missing data. Missingness mechanisms are classified into three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin 1976).

Assumptions about longitudinal data that include missing responses caused by dropouts are classified as follows:

- The data are said to be MCAR if the probability of a missing response is independent of its past, current, and future responses conditional on the covariates. That is, $P(r_{ij} = 0|\mathbf{Y}_i, \mathbf{X}_i) = P(r_{ij} = 0|\mathbf{X}_i)$.
- The data are said to be MAR if the probability of a missing response is independent of its current and future responses conditional on the observed past responses and the covariates. That is, $P(r_{ij} = 0|r_{ij-1} = 1, X_i, Y_i) = P(r_{ij} = 0|r_{ij-1} = 1, X_i, y_{i1}, \ldots, y_{ij-1})$. MAR is a weaker assumption than MCAR.
- The data are said to be MNAR if the probability of a missing response depends on the unobserved responses. MNAR is the most general and the most problematic missing-data scenario.

The GEE procedure implements two different weighted methods (observation-specific and subject-specific) of estimating the regression parameter $\boldsymbol{\beta}$ when dropouts occur. Both methods provide consistent estimates if the data are MAR. The weighted GEE methods are not supported for the multinomial distribution for polytomous responses.

Observation-Specific Weighted GEE Method

Suppose w_{ij} is the weight for y_{ij}, which is defined as the inverse probability of observing y_{ij}. In other words, $w_{ij} = P(r_{ij} = 1|X_i, Y_i)^{-1}$. Suppose W_i is a $T \times T$ diagonal matrix whose *j*th diagonal is $r_{ij} w_{ij}$. The responses for the *i*th subject are $\mathbf{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$. Consider the following weighted generalized estimating equations (Robins and Rotnitzky 1995; Preisser, Lohman, and Rathouz 2002):

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

Unlike the standard generalized estimating equations, the weighted generalized estimating equations are unbiased when the observations are appropriately weighted and lead to consistent estimates of $\boldsymbol{\beta}$.

The weights w_{ij} are often unknown in practice and are estimated by a logistic regression model under the MAR assumption. Specifically, suppose that $\lambda_{ij} = P(r_{ij} = 1|r_{ij-1} = 1, X_i, Y_i)$ denotes the probability of observing the response y_{ij} given its observed previous responses.

Under the MAR assumption,

$$\lambda_{ij} = P(r_{ij} = 1|r_{ij-1} = 1, X_i, Y_i) = P(r_{ij} = 1|r_{ij-1} = 1, X_i, Y_1, \ldots, Y_{j-1})$$

Using the observed data, λ_{ij} can be predicted from a logistic regression model,

$$\text{logit}\{\lambda_{ij}\} = z_{ij}\boldsymbol{\alpha}$$

where the z_{ij} are predictors that usually include the covariates x_{ij}, the past responses, and the indicators for visit times. The dropout process implies that the estimated probability of observing y_{ij} can be expressed as a cumulative product of conditional probabilities:

$$\hat{P}(r_{ij} = 1|X_i, Y_i) = \lambda_{i1}(\hat{\boldsymbol{\alpha}}) \times \lambda_{i2}(\hat{\boldsymbol{\alpha}}) \times \cdots \times \lambda_{ij}(\hat{\boldsymbol{\alpha}})$$

With the estimated weights $\hat{w}_{ij} = \hat{P}(r_{ij} = 1|X_i, Y_i)^{-1}$, the regression parameter $\boldsymbol{\beta}$ is estimated by solving the equation for $\mathbf{S}_{ow}(\boldsymbol{\beta})$.

The regression parameter $\boldsymbol{\beta}$ can be estimated by solving for $\mathbf{S}_{ow}(\boldsymbol{\beta})$ after plugging in the estimated weights. The fitting algorithm is described in the section “Fitting Algorithm for Weighted GEE” on page 3379.

Subject-Specific Weighted GEE Method

Unlike the observation-specific weighted method, which assigns an observation-specific weight to each observation, the subject-specific weighted method assigns a single weight to each subject. In other words, all the observations from a subject receive the same weight. Specifically, the subject-specific weighted method obtains the regression parameter estimates by solving the equations

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

where the responses for the *i*th subject are $\mathbf{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$ and the weight w_i for subject *i* is the inverse probability of a subject *i* dropping out at the observed time (Fitzmaurice, Molenberghs, and Lipsitz 1995; Preisser, Lohman, and Rathouz 2002). Note that the weight w_i is a scalar, in contrast to the weight matrix $\mathbf{W}_i$ that the observation-specific weighted GEE method uses.

The subject-specific weighted estimating equations are also unbiased when the subjects are appropriately weighted and lead to consistent estimates of the regression parameters $\boldsymbol{\beta}$.

The weight w_i is usually unknown in practice and needs to be estimated. Suppose subject *i* drops out at time $m_i = \sum_{j=1}^{T} r_{ij} + 1$. Assume that the first visit y_{i1} is always observed with $r_{i1} = 1$. Thus, the dropout times m_i range from 2 to *T+1*. Note that a dropout time of *T+1* indicates that subject *i* completes all the *T* visits and dropout does not occur.

The weight w_i is defined as follows: if subject *i* drops out before completing the last visit (that is, $m_i \le T$), then $w_i = P(r_{im_i} = 0, r_{im_i-1} = 1|X_i, Y_i)^{-1}$; otherwise, the subject completes all the T visits (that is, $m_i = T + 1$), and $w_i = P(r_{iT} = 1|X_i, Y_i)^{-1}$.

Similar to the process for the observation-specific weighted method, the dropout process for the subject-specific weighted method implies that subject-specific weights can be estimated as a cumulative product of conditional probabilities:

$$\hat{w}_i = P(r_{im_i} = 0, r_{im_i-1} = 1|X_i, Y_i)^{-1} = [\lambda_{i1}(\hat{\boldsymbol{\alpha}}) \times \cdots \times \lambda_{im_i-1}(\hat{\boldsymbol{\alpha}}) \times (1 - \lambda_{im_i}(\hat{\boldsymbol{\alpha}}))]^{-1}, \text{ if } m_i \le T$$
$$\hat{w}_i = P(r_{im_i-1} = 1|X_i, Y_i)^{-1} = [\lambda_{i1}(\hat{\boldsymbol{\alpha}}) \times \lambda_{i2}(\hat{\boldsymbol{\alpha}}) \times \cdots \times \lambda_{im_i-1}(\hat{\boldsymbol{\alpha}})]^{-1}, \text{ if } m_i = T + 1$$

Thus, the subject-specific weights $\hat{w}_i$ can be obtained after λ_{ij} is estimated by fitting a logistic regression to the data (r_{ij}, z_{ij}).

The regression parameter $\boldsymbol{\beta}$ from the subject-specific weighted GEE method can be estimated by solving for $\mathbf{S}_{sw}(\boldsymbol{\beta})$ after plugging in the estimated weights. The fitting algorithm is described in the section “Fitting

Algorithm for Weighted GEE" on page 3379. The subject-specific weighting scheme was originally developed for computational convenience. Preisser, Lohman, and Rathouz (2002) showed that the observation-level weighted GEE method produces more efficient estimates than the cluster-level weighted GEE method for incomplete longitudinal binary data.

Fitting Algorithm for Weighted GEE

The following fitting algorithm fits marginal models by using the observation-specific or the subject-specific weighted GEE method when the dropout process is missing at random:

1. Fit a logistic regression to the data (r_{ij}, z_{ij}) to obtain an estimate of $\boldsymbol{\alpha}$ and estimate the weights.

2. Compute an initial estimate of $\boldsymbol{\beta}$ by using an ordinary generalized linear model, assuming independence of the responses.

3. Compute the working correlation matrix $\mathbf{R}$ based on the standardized residuals, the current estimate of $\boldsymbol{\beta}$, and the specified structure of $\mathbf{R}$.

4. Compute the estimated covariance matrix:

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

5. Update $\hat{\boldsymbol{\beta}}$:

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

   where $\mathbf{Y}_i$, $\boldsymbol{\mu}_i$, $\mathbf{V}_i$, and $\mathbf{W}_i$ are as follows:

   - For the observation-specific weighted method, $\mathbf{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$; $\boldsymbol{\mu}_i$ and $\mathbf{V}_i$ are its corresponding mean vector and working covariance matrix, respectively; and $\mathbf{W}_i$ is a $T \times T$ diagonal matrix whose jth diagonal is $r_{ij}\hat{w}_{ij}$.
   - For the subject-specific weighted method, $\mathbf{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$; $\boldsymbol{\mu}_i$ and $\mathbf{V}_i$ are its corresponding mean vector and working covariance matrix, respectively; and $\mathbf{W}_i$ is a $n_i \times n_i$ diagonal matrix whose jth diagonal is $\hat{w}_i$.

6. Repeat steps 3–5 until convergence.

Note that you can use the WEIGHT statement in the GENMOD procedure to perform a two-stage strategy that is often used in practice to obtain the weighted GEE estimates. You fit a logistic regression to the data (r_{ij}, z_{ij}) to obtain the weights as described in the preceding steps. Then you estimate $\boldsymbol{\beta}$ by specifying the estimated weights in the WEIGHT statement in PROC GENMOD for the GEE analysis. For the subject-specific weighted GEE method, this approach is appropriate for any working correlation structure. However, for the observation-specific weighted method, this approach is appropriate only for the independent working correlation structure.

The two-stage approach results in standard errors that are larger than those that are produced by using the MISSMODEL statement in the GEE procedure (because PROC GENMOD treats the weights as fixed and known). Thus, the two-stage approach that uses PROC GENMOD results in conservative inference (Fitzmaurice, Laird, and Ware 2011). The GEE procedure computes the parameter estimate covariances as described in (Fitzmaurice, Laird, and Ware 2011) and Preisser, Lohman, and Rathouz (2002).

Missing Data

Suppose that each subject in a longitudinal study is measured at T times. In other words, for the ith subject you measure T responses $(y_{i1}, y_{i2}, \ldots, y_{iT})$ and T corresponding covariates $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, , \ldots, x_{iT})$.

By default, the GEE procedure handles missing data in the same manner as the standard GEE method in the GENMOD procedure. The working correlation matrix is estimated from data that contain both intermittent and dropout types of missing values by using the all-available-pairs method, in which all nonmissing pairs of data are used in the moment estimators. The resulting covariances and standard errors are valid under the missing completely at random (MCAR) assumption. For more information, see the section “Missing Data” on page 3517 in Chapter 48, “The GENMOD Procedure.”

When you specify the MISSMODEL statement in the GEE procedure to use the weighted GEE method to analyze the data, the procedure uses observations that have missing values in the response, provided that the missing values for all subjects are caused by dropouts. If the missing values are intermittent for any of the subjects, then the weighted GEE method does not apply and the procedure terminates.

For the observation-specific weighted GEE method, the covariates for all the observations for a subject must be observed, regardless of whether the response is missing. For each subject, the input data set must provide T observations.

For the subject-specific weighted GEE method, the covariates for a subject who drops out at time k must be observed for the observations up to and including time k. The input data set must provide at least k observations for this subject. The covariates must be observed for all observations on a subject who completes the study, and the input data set must provide T observations for this subject.

For more information about how weighted GEE methods handle missing values, see Fitzmaurice, Laird, and Ware (2011) and Preisser, Lohman, and Rathouz (2002).

Type 3 Analysis

A Type 3 analysis is similar to the Type 3 sums of squares used in PROC GLM, except that generalized score tests for Type 3 contrasts instead of Type 3 sums of squares are computed. Briefly, a Type 3 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 3 contrast for a main effect is equal to 0 is intended to test the significance of the main effect in the presence of interactions. For more information about Type 3 estimable functions, see Chapter 50, “The GLM Procedure,” and Chapter 15, “The Four Types of Estimable Functions.” Also see Littell, Freund, and Spector (1991).

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

The generalized score statistic is

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

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

A Type 3 analysis can consume considerable computation time because 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 $\mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$ is defined by

$$S = (\mathbf{L}'\hat{\boldsymbol{\beta}})'(\mathbf{L}'\boldsymbol{\Sigma}_e\mathbf{L})^{-1}(\mathbf{L}'\hat{\boldsymbol{\beta}})$$

where $\mathbf{L}$ is the contrast matrix, $\boldsymbol{\beta}$ are the GEE parameter estimates, and $\boldsymbol{\Sigma}_e$ is the empirical covariance estimate. The asymptotic distribution of S is chi-square with r degrees of freedom, where r is the rank of $\mathbf{L}$.

The results of this type of analysis do not depend on the order in which the terms are specified in the MODEL statement. Type 3 analyses that use score statistics are not supported for nominal response data or weighted GEE methods. Type 3 analyses can be conducted using the Wald statistics for all the models that the GEE procedure supports.

ODS Table Names

PROC GEE assigns a name to each table that it creates. You can use these names to refer to the table when you use the Output Delivery System (ODS) to select tables and create output data sets. Table 47.14 lists these names. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

Table 47.14 ODS Tables Produced BY PROC GEE

ODS Table Name	**Description**	**Statement**	**Option**
ClassLevels	Classification variable levels	CLASS	Default
Coef	Coefficients for LS-means	LSMEANS	E
Diffs	Differences of LS-means	LSMEANS	DIFF
Estimates	Estimates of contrasts	ESTIMATE	Default
GEEEmpPEst	Parameter estimates with empirical standard errors	REPEATED	Default
GEEExchCorr	Exchangeable working correlation value	REPEATED	TYPE=EXCH
GEEFitCriteria	QIC fit criteria	REPEATED	Default
GEELogORInfor	GEE log odds ratio model information	REPEATED	LOGOR=
GEEModInfo	GEE model information	REPEATED	Default
GEEModPEst	Parameter estimates with model-based standard errors	REPEATED	MODELSE
GEENCorr	Model-based correlation matrix	REPEATED	MCORRB
GEENCov	Model-based covariance matrix	REPEATED	MCOVB
GEERCorr	Empirical correlation matrix	REPEATED	ECORRB
GEERCov	Empirical covariance matrix	REPEATED	ECOVB
GEEWCorr	GEE working correlation matrix	REPEATED	CORRW
LSMeans	LS-means	LSMEANS	Default

Table 47.14 *continued*

ODS Table Name	Description	Statement	Option
LSMLines	Lines display for LS-means	LSMEANS	LINES
MissModelPEst	Parameter estimates for the missingness model	MISSMODEL	Default
MissPattern	Frequency counts for dropout times	MISSMODEL	Default
ModelInfo	Model information	MODEL	Default
NObs	Number of observations summary		Default
ParmInfo	Parameter indices	REPEATED	MCORRB, MCOVB, ECORRB, ECOVB
ResponseProfile	Frequency counts for binary models	MODEL	DIST=BINOMIAL
Type3	Type 3 tests	MODEL	TYPE3

ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, "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 623 in Chapter 21, "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 622 in Chapter 21, "Statistical Graphics Using ODS."

ODS Graph Names

PROC GEE assigns a name to each graph it creates using ODS. You can use these names to refer to the graphs when you use ODS. Table 47.15 lists the names.

To request these graphs, ODS Graphics must be enabled and you must specify the statement and option that are indicated in Table 47.15.

Table 47.15 Graphs Produced by PROC GEE

ODS Graph Name	Description	Statement	Option
Histogram	Histogram of predicted weights from the missingness model	PROC	PLOTS=

Examples: GEE Procedure

The following examples illustrate some of the capabilities of the GEE procedure. These examples are not intended to represent definitive analyses of the data sets that are presented here.

Example 47.1: Comparison of the Marginal and Random Effect Models for Binary Data

A clinical trial (Stokes, Davis, and Koch 2012) was conducted to compare two treatments for a respiratory illness. Patients in each of two centers were randomly assigned to two groups: one group received the active treatment and one group received a placebo.

During treatment, respiratory status was determined for each of four visits and is represented by the variable Outcome (coded here as 0 = poor, 1 = good). The variables Center, Treatment, Sex, and Baseline (baseline respiratory status) are classification variables that have two levels. The variable Age (age at time of entry into the study) is a continuous variable.

All 111 patients completed the study. That is, there are no missing data for responses or covariates. The following statements create the data set Resp:

```
data Resp;
   input Center ID Treatment $ Sex $ Age Baseline Visit1-Visit4;
   datalines;
1   1 P M 46 0 0 0 0 0
1   2 P M 28 0 0 0 0 0
1   3 A M 23 1 1 1 1 1
1   4 P M 44 1 1 1 1 0
1   5 P F 13 1 1 1 1 1
1   6 A M 34 0 0 0 0 0

   ... more lines ...

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

data Resp;
   set Resp;
   Visit=1;  Outcome=Visit1;  output;
   Visit=2;  Outcome=Visit2;  output;
   Visit=3;  Outcome=Visit3;  output;
   Visit=4;  Outcome=Visit4;  output;
run;
```

Suppose y_{ij} represents the respiratory status of patient i at the jth visit, $j = 1, \ldots, 4$, and $\mu_{ij} = \mathrm{E}(y_{ij})$ represents the mean of the respiratory status. Logistic regression is commonly used to analyze binary response data. 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.

The following SAS statements perform the GEE model fit:

```
proc gee data=Resp descend;
   class ID Treatment Center Sex Baseline;
   model Outcome=Treatment Center Sex Age Baseline /
         dist=bin link=logit;
   repeated subject=ID(Center) / corr=exch corrw;
run;
```

Both the MODEL statement and the REPEATED statement are required.

In the MODEL statement, you use the DIST=BIN and LINK=LOGIT options to specify a logistic regression, and you specify Outcome as the response variable and Treatment, Center, Sex, Age, and Baseline as the explanatory variables. The DESCEND option in the PROC GEE statement requests that the probability that Outcome = 1 be modeled. If the DESCEND option had not been specified, the probability that Outcome = 0 would be modeled by default.

You use the REPEATED statement to specify the subject and the correlation structure of the responses. The SUBJECT=ID(CENTER) option specifies that the observations in any single cluster are uniquely identified by Center and ID. An equivalent specification is SUBJECT=ID*CENTER. Because 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 could be specified. The option TYPE=EXCH specifies the exchangeable working correlation structure.

The "Model Information" table displayed in Output 47.1.1 provides information about the specified logistic regression model and the input data set.

Output 47.1.1 Model Information

The GEE Procedure

```
     Model Information
Data Set            WORK.RESP
Distribution         Binomial
Link Function           Logit
Dependent Variable    Outcome
```

General information about the GEE analysis is displayed in Output 47.1.2, and model fit criteria for the model are displayed in Output 47.1.3.

Output 47.1.2 Model Fitting Information

GEE Model Information	
Correlation Structure	Exchangeable
Subject Effect	ID(Center) (111 levels)
Number of Clusters	111
Correlation Matrix Dimension	4
Maximum Cluster Size	4
Minimum Cluster Size	4

Output 47.1.3 Model Fitting Information

GEE Fit Criteria	
QIC	512.5723
QICu	499.4873

The results of GEE model fitting are displayed in Output 47.1.4. 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 that is based on model-based standard error estimates.

Output 47.1.4 Results of Model Fitting

Parameter Estimates for Response Model with Empirical Standard Error Estimates

Parameter		**Estimate**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
Intercept		1.6391	0.5247	0.6107	2.6675	3.12	0.0018
Treatment	A	1.2654	0.3467	0.5859	1.9448	3.65	0.0003
Treatment	P	0.0000	0.0000	0.0000	0.0000	.	.
Center	1	-0.6495	0.3532	-1.3418	0.0428	-1.84	0.0660
Center	2	0.0000	0.0000	0.0000	0.0000	.	.
Sex	F	0.1368	0.4402	-0.7261	0.9996	0.31	0.7560
Sex	M	0.0000	0.0000	0.0000	0.0000	.	.
Age		-0.0188	0.0130	-0.0442	0.0067	-1.45	0.1480
Baseline	0	-1.8457	0.3460	-2.5238	-1.1676	-5.33	<.0001
Baseline	1	0.0000	0.0000	0.0000	0.0000	.	.

Treatment and Baseline appear to be strongly influential, and Center might be marginally significant.

For comparison, a generalized linear mixed model is fitted to the data set to obtain subject-specific effects. Specifically, consider the logistic regression model,

$$\mathrm{logit}(\mathrm{E}(y_{ij}|b_i)) = \mathbf{x}_{ij}'\boldsymbol{\beta}^* + b_i$$

where the random effect b_i is normally distributed with zero mean and variance, $\mathrm{Var}(b_i) = \sigma_b^2$.

The following statements use the GLIMMIX procedure to fit a generalized linear mixed model:

```
proc glimmix data=Resp;
   class ID Treatment Center Sex Baseline;
   model Outcome (desc)=Treatment Center Sex Age Baseline /
         dist=binary solution;
   random ID(Center);
run;
```

Output 47.1.5 displays the parameter estimates for the fixed effects in the generalized linear mixed model.

Output 47.1.5 Parameter Estimates

The GLIMMIX Procedure

Solutions for Fixed Effects

Effect	Treatment	Sex	Center	Baseline	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept					1.7936	0.6292	105	2.85	0.0053
Treatment	A				1.4758	0.3898	333	3.79	0.0002
Treatment	P				0	.	.	.	.
Center			1		-0.7201	0.4051	105	-1.78	0.0784
Center			2		0	.	.	.	.
Sex		F			0.1732	0.5034	333	0.34	0.7310
Sex		M			0	.	.	.	.
Age					-0.02011	0.01507	333	-1.33	0.1831
Baseline				0	-2.1343	0.3971	333	-5.38	<.0001
Baseline				1	0	.	.	.	.

From Output 47.1.4 and Output 47.1.5, you can see that the parameter estimates from the marginal model and the mixed-effects model differ. For example, the estimated treatment effects are 1.2654 and 1.4758 from the marginal model and the mixed-effects model, respectively.

The interpretation of the model effects in the marginal and random models differs. For example, the estimated treatment effect from the marginal model indicates that, on average, the odds of a good response for the patients is $e^{1.2654} = 3.5$ times higher when they receive the active treatment versus the placebo. The estimated treatment effect from the generalized linear mixed model indicates that an individual patient's odds of a good response is $e^{1.4758} = 4.4$ times higher when the patient receives the active treatment versus the placebo.

The choice of the marginal model or a subject-specific model often depends on the goal of your analysis: whether you are interested in population-averaged effects or subject-specific effects. For more information, see Diggle et al. (2002); Fitzmaurice, Laird, and Ware (2011).

Example 47.2: Log-Linear Model for Count Data

The following example demonstrates how you can fit a GEE model to count data. The data are analyzed by Diggle, Liang, and Zeger (1994). The response is the number of epileptic seizures, which was measured at the end of each of eight two-week treatment periods over sixteen weeks. The first eight weeks were the baseline period (during which no treatment was given), and the second eight weeks were the treatment period, during which patients received either a placebo or the drug progabide. The question of scientific interest is whether progabide is effective in reducing the rate of epileptic seizures.

The following DATA step creates the data set Seizure:

```
data Seizure;
   input ID Count Visit  Trt Age Weeks;
   datalines;
104 11 0 0 31 8
104 5 1 0 31 2
104 3 2 0 31 2
104 3 3 0 31 2
104 3 4 0 31 2
106 11 0 0 30 8

   ... more lines ...

236 12 0 1 37 8
236 1 1 1 37 2
236 4 2 1 37 2
236 3 3 1 37 2
236 2 4 1 37 2
;
```

The following DATA step creates 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, which was done in the Diggle et al. (2002) analysis:

```
data Seizure;
   set Seizure;
   if ID ne 207;
   if Visit = 0 then do;
      X1=0;
      Ltime = log(8);
   end;
   else do;
      X1=1;
      Ltime=log(2);
   end;
run;
```

Poisson regression is commonly used to model count data. In this example, the log-linear Poisson model is specified by $V(\mu) = \mu$ (the Poisson variance function) and a log link function,

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

where

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

Because the visits represent repeated measurements, the responses from the same individual are correlated and inferences need to take this into account. The correlations between the counts are modeled as $r_{ij} = \alpha$, $i \neq j$ (exchangeable correlations).

In this model, the regression parameters are interpreted in terms of the log seizure rate that is displayed in Table 47.16.

Table 47.16 Interpretation of Regression Parameters

Treatment	**Visit**	$\log(E(Y_{ij})/t_{ij})$
Placebo	Baseline	β_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 β_1 for the placebo group and $\beta_1 + \beta_3$ for the progabide group. A value of $\beta_3 < 0$ indicates a reduction in the seizure rate.

The following statements perform the analysis:

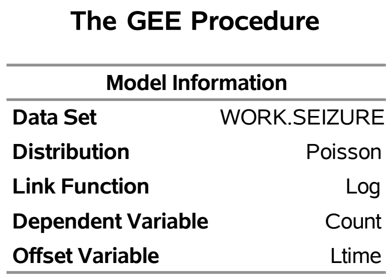
```
proc gee data = Seizure;
   class ID Visit;
   model Count = X1 Trt X1*Trt / dist=poisson link=log offset= Ltime;
   repeated subject = ID / within = Visit type=unstr covb corrw;
run;
```

In the MODEL statement, Count is the response variable, and X1, Trt, and the interaction X1*Trt are the explanatory variables. You request Poisson regression with the DIST=POISSON and the LINK=LOG options. The offset variable is often used in Poisson regression to account for different exposures. In this case, the OFFSET= option specifies Ltime as the offset variable representing different time intervals.

In the REPEATED statement, the SUBJECT= option indicates that the variable ID identifies the observations from a single cluster, and the TYPE=UNSTR option specifies the unstructured working correlation structure. The CORRW option requests that the working correlation matrix be displayed.

The "Model Information" table that is displayed in Output 47.2.1 provides information about the specified model and the input data set.

Output 47.2.1 Model Information



The GEE Procedure

Model Information	
Data Set	WORK.SEIZURE
Distribution	Poisson
Link Function	Log
Dependent Variable	Count
Offset Variable	Ltime

Output 47.2.2 displays general information about the GEE model analysis.

Output 47.2.2 GEE Model Information

GEE Model Information	
Correlation Structure	Unstructured
Within-Subject Effect	Visit (5 levels)
Subject Effect	ID (58 levels)
Number of Clusters	58
Correlation Matrix Dimension	5
Maximum Cluster Size	5
Minimum Cluster Size	5

Output 47.2.3 displays the parameter estimate covariance matrices, which are requested by the COVB option. Both model-based and empirical covariances are produced.

Output 47.2.3 Covariance Matrices of Parameter Estimate

Covariance Matrix (Model-Based)

	Prm1	Prm2	Prm3	Prm4
Prm1	0.01210	0.004902	-0.01210	-0.004902
Prm2	0.004902	0.006660	-0.004902	-0.006660
Prm3	-0.01210	-0.004902	0.02461	0.01299
Prm4	-0.004902	-0.006660	0.01299	0.01852

Covariance Matrix (Empirical)

	Prm1	Prm2	Prm3	Prm4
Prm1	0.02597	-0.003069	-0.02597	0.003069
Prm2	-0.003069	0.008597	0.003069	-0.008597
Prm3	-0.02597	0.003069	0.03841	-0.006196
Prm4	0.003069	-0.008597	-0.006196	0.02237

The exchangeable working correlation matrix is displayed in Output 47.2.4. It shows that there are noticeable correlations among the respective visits.

Output 47.2.4 Working Correlation Matrix

Working Correlation Matrix

	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5
Obs 1	1.0000	0.7920	0.7190	0.8111	0.6582
Obs 2	0.7920	1.0000	0.4859	0.6552	0.4566
Obs 3	0.7190	0.4859	1.0000	0.6988	0.4171
Obs 4	0.8111	0.6552	0.6988	1.0000	0.6464
Obs 5	0.6582	0.4566	0.4171	0.6464	1.0000

The parameter estimates table, shown in Output 47.2.5, 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.

Output 47.2.5 Parameter Estimates Table

Parameter Estimates for Response Model
with Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	1.3309	0.1612	1.0151	1.6468	8.26	<.0001
X1	0.1128	0.0927	-0.0689	0.2945	1.22	0.2237
Trt	-0.1034	0.1960	-0.4875	0.2807	-0.53	0.5978
X1*Trt	-0.3162	0.1496	-0.6093	-0.0231	-2.11	0.0345

The estimate of β_3 is –0.3162, which indicates that progabide is effective in reducing the rate of epileptic seizures.

Model fit criteria for the model are displayed in Output 47.2.6. These criteria are used in selecting regression models and working correlations.

Output 47.2.6 Model Fit Criteria

GEE Fit Criteria	
QIC	-1036.2837
QICu	-1041.8041

Example 47.3: Weighted GEE for Longitudinal Data That Have Missing Values

This example shows how you can use the GEE procedure to analyze longitudinal data that contain missing values. The data set is taken from a longitudinal study of women who used contraception during one year (Fitzmaurice, Laird, and Ware 2011). In this study, 1,151 women were randomly assigned to one of two treatments: 100 mg or 150 mg of depot medroxyprogesterone acetate (DMPA) at baseline and at three-month intervals. The response variable indicates their amenorrhea status during the four three-month intervals. The question of interest is whether the treatment has an effect on the rate of the amenorrhea over time. The example follows the analysis done by Fitzmaurice, Laird, and Ware (2011).

The following statements create the data set Amenorrhea:

```
data Amenorrhea;
   input ID Dose Time Y@@;
   datalines;
   1     0       1     0
   1     0       2     .
   1     0       3     .
   1     0       4     .

   ... more lines ...

1150     1       4     1
1151     1       1     1
1151     1       2     1
1151     1       3     1
1151     1       4     1
;
```

The variables in the data are as follows:

- ID: patient's ID
- Y: indicator of amenorrhea status (1 for amenorrhea; 0 otherwise)
- Time: four consecutive three-month intervals with values 1, 2, 3, and 4
- Dose: 0 for treatment with 100 mg injection; 1 for treatment with 150 mg injection

To prepare for the analysis, two additional variables are created:

- Prevy: the patient's amenorrhea status in the previous three-month interval. For the baseline visit, this is set to an arbitrary nonmissing value (0 here). In this release of PROC GEE, this arbitrary value must be nonmissing and valid for the response variable—for example, it should be 0 or 1 for a binary response—but it does not otherwise affect the results.
- Ctime: a copy of Time, which you can include in the marginal model as a continuous effect and also in the missingness model as a classification effect

The following statements add these two variables to the data set:

```
data Amenorrhea;
   set Amenorrhea;
   by ID;
   Prevy=lag(Y);
   if first.id then Prevy=0;
   Time=Time-1;
   Ctime=Time;
run;
```

Suppose y_{ij} denotes the amenorrhea status of woman i at the jth visit, $j = 1, \ldots, 4$, and suppose $\mu_{ij} = \mathrm{P}(y_{ij} = 1)$ denotes the average rate of high dosage. To explore whether the treatment has an effect on the rate of amenorrhea over time, consider the following marginal model:

$$\mathrm{logit}(\mu_{ij}) = \beta_0 + \beta_1 \mathrm{time}_{ij} + \beta_2 \mathrm{time}_{ij}^2 + \beta_3 \mathrm{dose}_i + \beta_4 \mathrm{dose}_i \times \mathrm{time} + \beta_5 \mathrm{dose}_i \times \mathrm{time}^2$$

Of the 1,151 women in this study, 576 are from the low-dose group, and 575 are from the high-dose group. For the low-dose group, 62.67% of the women completed the trial; for the high-dose group, 61.39% of the women completed this trial. Thus, both groups have substantial dropouts.

To obtain the weights for the weighted GEE analysis, consider the following logistic regression model for missingness:

$$\begin{aligned}\mathrm{logit}\, p(r_{ij} = 1 | r_{ij-1} = 1, \mathrm{dose}_i, \mathrm{ctime}_{ij}, y_{ij-1}) =& \alpha_0 + \alpha_1 I(\mathrm{ctime}_{ij} = 1) + \alpha_2 I(\mathrm{ctime}_{ij} = 2) \\ & + \alpha_3 \mathrm{dose}_i + \alpha_4 y_{ij-1} + \alpha_5 \mathrm{dose}_i \times y_{ij-1}\end{aligned}$$

The following statements use the observation-specific weighted GEE method and the specified response and missingness models to analyze the data:

```
ods graphics on;
proc gee data=Amenorrhea desc plots=histogram;
   class ID Ctime;
   missmodel Ctime Prevy Dose Dose*Prevy / type=obslevel;
   model Y = Time Dose Time*Time Dose*Time Dose*Time*Time / dist=bin;
   repeated subject=ID / within=Ctime corr=cs;
run;
```

The MODEL statement specifies logistic regression and the model effects. The DESCEND option in the PROC GEE statement models the probability that Y = 1.

The REPEATED statement requests GEE analysis. The SUBJECT=ID option specifies that observations from the same subject are identified by ID. The TYPE=CS option specifies the compound symmetric working correlation structure.

The MISSMODEL statement requests the weighted GEE analysis. It specifies the logistic regression model for missingness. Note that no response variable is needed in weighted GEE analysis to specify a missingness model because the response is completely determined by the response variable in the MODEL statement. Without the MISSMODEL statement, PROC GEE would use the standard GEE approach, the same as provided by PROC GENMOD. The TYPE=OBSLEVEL option requests observation-specific weights.

Output 47.3.1 shows the parameter estimates for the missingness model. The estimate of α_4 is –0.4514 with a p-value of 0.0053, which suggests that the possibility that a participant will drop out is related to her previous amenorrhea status. This suggests that the assumption of MAR is more appropriate than that of MCAR.

Output 47.3.1 Parameter Estimates for the Missingness Model

Parameter Estimates for Missingness Model

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		2.3967	0.1438	2.1149	2.6785	16.67	<.0001
Ctime	0	0.0000	0.0000	0.0000	0.0000	.	.
Ctime	1	-0.7286	0.1439	-1.0106	-0.4466	-5.06	<.0001
Ctime	2	-0.5919	0.1469	-0.8798	-0.3040	-4.03	<.0001
Ctime	3	0.0000	0.0000	0.0000	0.0000	.	.
Prevy		-0.4514	0.1619	-0.7687	-0.1341	-2.79	0.0053
Dose		0.0680	0.1313	-0.1893	0.3253	0.52	0.6046
Prevy*Dose		-0.2381	0.2196	-0.6685	0.1923	-1.08	0.2782

The classification variable Ctime has two levels whose estimates are equal to zero. One is the reference level Ctime = 3. The first level, Ctime = 0, also has an estimate of zero, because the first visit is always observed and the first level is never used in estimating the weights in the missing model.

Output 47.3.2 displays the results of the weighted GEE analysis.

Output 47.3.2 Parameter Estimates for Amenorrhea Data Analysis Using Weighted GEE

The GEE Procedure

Parameter Estimates for Response Model with Empirical Standard Error Estimates

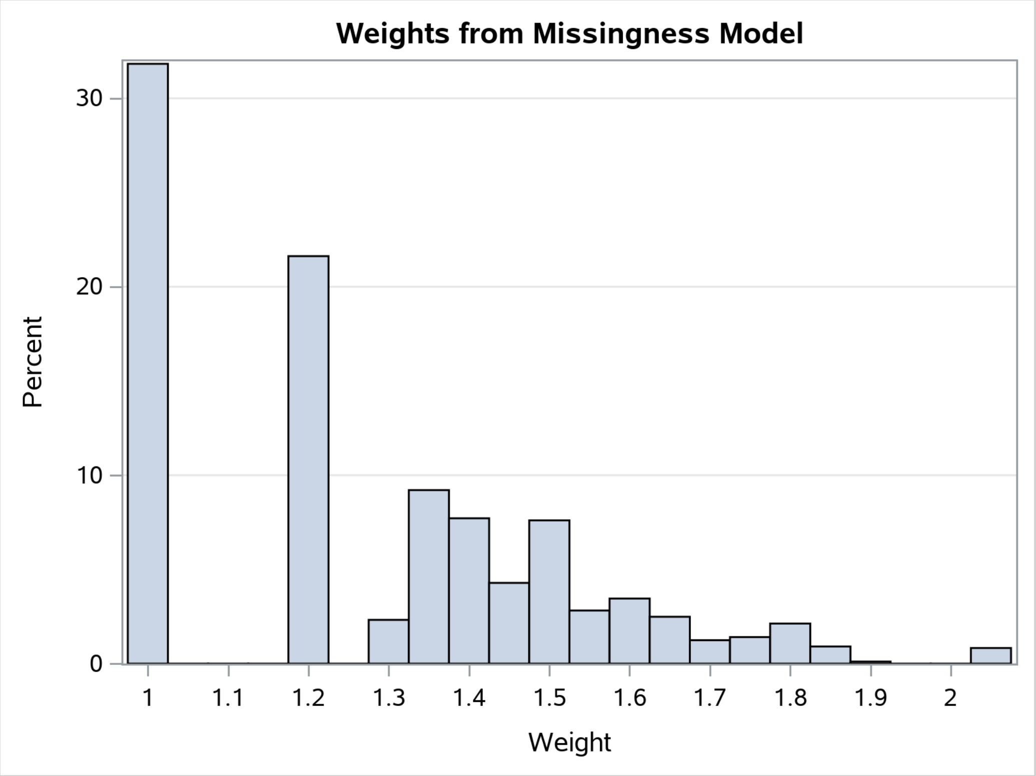
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-1.4965	0.1072	-1.7067	-1.2863	-13.95	<.0001
Time	0.5379	0.1334	0.2764	0.7994	4.03	<.0001
Dose	0.1061	0.1491	-0.1861	0.3983	0.71	0.4767
Time*Time	-0.0037	0.0405	-0.0831	0.0757	-0.09	0.9275
Dose*Time	0.4092	0.1903	0.0362	0.7823	2.15	0.0315
Dose*Time*Time	-0.1264	0.0577	-0.2395	-0.0134	-2.19	0.0284

The estimate of β_4 (the parameter estimate for the Dose*Time interaction) is 0.4092, which indicates that the change of amenorrhea rate over time depends on the dose of DMPA. Specifically, for women in the low-dose group, the amenorrhea rates μ_{ij} at the four consecutive time intervals are 0.1830, 0.2764, 0.3928, and 0.5210 and for women in the high-dose group, the amenorrhea rate are 0.1997, 0.3609, 0.4963, and 0.5701. In other words, the amenorrhea rate increases over time for both treatments, and the rates of increase are slightly different.

You can request subject-level weights by specifying the TYPE=SUBLEVEL option. The results (not shown here) from the subject-level weighted method are similar to the results from the observation-level weighted method. Both of the weighted GEE methods provide unbiased regression parameter estimates if the missingness model is specified correctly. Preisser, Lohman, and Rathouz (2002) note that the observation-level weighted GEE produces more efficient estimates than the cluster-level weighted GEE produces for incomplete longitudinal binary data.

Large weights can have impacts on the parameter estimates. Consequently, it is recommended that you check the distribution of the estimated weights. If there are large weights, you might consider trimming them by specifying the MAXWEIGHT= option in the MISSMODEL statement. Output 47.3.3 shows that the estimated weights in this example range between 1 and 2.1, so no trimming is needed.

Output 47.3.3 Histogram of Estimated Weights



Example 47.4: GEE for Binary Data with Logit Link Function

Because the respiratory data in Example 47.1 are binary, you can use the alternating logistic regression (ALR) method and model associations by using the log odds ratios instead of working correlations. This example fits a “fully parameterized cluster” model for the log odds ratio. 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 47.1 but use a regression model for the log odds ratios instead of a working correlation. LOGOR=FULLCLUST specifies a fully parameterized log odds ratio model.

```
proc gee data=Resp descend;
   class ID Treatment Center Sex Baseline;
   model Outcome=Treatment Center Sex Age Baseline / dist=bin;
   repeated  subject=ID(Center) / logor=fullclust;
run;
```

The results of fitting the model are displayed in Output 47.4.1.

Output 47.4.1 Results of ALR Model Fitting

The GEE Procedure

Parameter Estimates for Response Model with Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		1.6001	0.5128	0.5950	2.6052	3.12	0.0018
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	1	-0.6287	0.3486	-1.3119	0.0545	-1.80	0.0713
Center	2	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	0	-1.8980	0.3404	-2.5652	-1.2308	-5.58	<.0001
Baseline	1	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

The parameters Alpha1 through Alpha6 estimate the log odds ratio for each unique within-cluster pair. The correspondence between the log odds ratio parameters and within-cluster pairs is displayed in Output 47.4.2.

Output 47.4.2 Log Odds Ratio Parameters

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)

Model goodness-of-fit criteria are shown in Output 47.4.3.

Output 47.4.3 ALR Model Fit Criteria

GEE Fit Criteria

QIC	511.8589
QICu	499.6516

The QIC for the ALR model shown in Output 47.4.3 is 511.86, whereas the QIC for the unstructured working correlation model shown in Output 47.1.3 is 512.34, indicating that the ALR model has a slightly better fit.

You can fit the same model by fully specifying the **z** matrix; for the definition of the **z** matrix, see the section "Specifying Log Odds Ratio Models" on page 3374. The following statements create a data set that contains 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 47.4.4 displays the full **z** matrix for the first two clusters. The **z** matrix is identical for all clusters in this example.

Output 47.4.4 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 gee data=Resp descend;
   class ID Treatment Center Sex Baseline;
   model Outcome=Treatment Center Sex Age Baseline / dist=bin;
   repeated  subject=ID(Center) / logor=zfull
                                  zdata=zin
                                  zrow =(z1-z6)
                                  ypair=(y1 y2);
run;
```

Example 47.5: Alternating Logistic Regression for Ordinal Multinomial Data

This example illustrates how you use the GEE procedure and alternating logistic regression (ALR) to analyze ordinal multinomial data. A clinical trial was conducted to evaluate the effectiveness of the drug auranofin for treating arthritis (Lipsitz, Kim, and Zhao 1994). Patients were assigned to one of two groups: one group was treated with auranofin, and the other group received a placebo. The treatment that a patient received is recorded in the variable Treatment (coded here as 1 = auranofin and 0 = placebo).

The response was self-assessment of arthritis recorded at one-, three-, and five-month follow-up visits. The responses are recorded in the Rating variable and are coded as 1 = very poor, 2 = poor, 3 = fair, 4 = good, and 5 = very good. This coding of Rating is finer than the coding in Lipsitz, Kim, and Zhao (1994), where only three levels were used. The visit numbers are recorded in the classification variable Visit, whose value is 1, 3, or 5. An initial self-assessment that uses the same coding as Rating is recorded in the variable Baseline. The variable Age records the participants' ages (in years) at the baseline visit and is treated as a continuous variable. One participant missed all visits and is not considered. There are an additional 15 missed visits from eight participants who dropped out, and there are four participants who missed a single visit. A weighted GEE is not used because the GEE procedure in SAS/STAT 14.1 does not support the weighted GEE method for the multinomial distribution.

The following DATA step creates the data set Arthritis:

```
data Arthritis;
   input ID Rating Sex Age Treatment Baseline Visit;
   datalines;
1 4 2 54 2 2 1
1 5 2 54 2 2 3
1 5 2 54 2 2 5
2 4 1 41 1 3 1
2 4 1 41 1 3 3
2 4 1 41 1 3 5

   ... more lines ...

301 2 2 64 1 2 5
302 2 2 55 1 2 1
302 3 2 55 1 2 3
302 3 2 55 1 2 5
;
```

The following SAS statements use PROC GEE to fit a model that has a fully exchangeable working correlation structure:

```
proc gee data=Arthritis;
   class Sex ID Treatment Baseline Visit;
   model Rating= Visit Treatment Baseline / dist=multinomial;
   repeated subject=ID / within=Visit logor=exch;
run;
```

You specify LOGOR=EXCH in the REPEATED statement to select the ALR method that has a fully exchangeable model for the log odds ratio. The results of the ALR model fitting are displayed in Output 47.5.1.

Output 47.5.1 Parameter Estimates for Arthritis Data Using ALR

The GEE Procedure

Parameter Estimates for Response Model
with Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept1		-6.7502	0.4267	-7.5865	-5.9138	-15.82	<.0001
Intercept2		-4.6310	0.3968	-5.4087	-3.8533	-11.67	<.0001
Intercept3		-2.6735	0.3749	-3.4083	-1.9387	-7.13	<.0001
Intercept4		-0.3838	0.3710	-1.1109	0.3433	-1.03	0.3008
Visit	1	0.3740	0.1148	0.1489	0.5991	3.26	0.0011
Visit	3	0.3641	0.1116	0.1455	0.5828	3.26	0.0011
Visit	5	0.0000	0.0000	0.0000	0.0000	.	.
Treatment	1	0.5552	0.1673	0.2273	0.8830	3.32	0.0009
Treatment	2	0.0000	0.0000	0.0000	0.0000	.	.
Baseline	1	3.9457	0.5352	2.8969	4.9946	7.37	<.0001
Baseline	2	3.3052	0.4268	2.4686	4.1418	7.74	<.0001
Baseline	3	2.7483	0.3790	2.0054	3.4911	7.25	<.0001
Baseline	4	1.4013	0.4132	0.5914	2.2113	3.39	0.0007
Baseline	5	0.0000	0.0000	0.0000	0.0000	.	.
Alpha1		1.6447	0.1693	1.3130	1.9764	9.72	<.0001

The parameter Alpha1, which is used to estimate the log odds ratio, is included in Output 47.5.1.

To fit the ALR model, each response is coded as a vector of binary variables and the log odds ratio models the association between pairs of responses. For more information about the log odds ratio and the ALR method for ordinal multinomial data, see the section "ALR for Ordinal Multinomial Data" on page 3373. The ALR model fit criteria are shown in Output 47.5.2.

Output 47.5.2 ALR Model Fit Criteria

GEE Fit Criteria	
QIC	2241.9540
QICu	2259.8575

For comparison, the following SAS statements use PROC GEE to fit the same marginal model by using an independent working correlation structure:

```
proc gee data=Arthritis;
   class Sex ID Treatment Baseline Visit;
   model Rating= Visit Treatment Baseline / dist=multinomial;
   repeated subject=ID / within=Visit;
run;
```

When the data have multinomial responses, the independent working correlation structure is the only structure supported for ordinary GEEs. In Output 47.5.1 and Output 47.5.3, you can see slight differences in the parameter estimates between the model that you fit by using ALR and the model that you fit by using an independent working correlation structure.

Output 47.5.3 Parameter Estimates for Arthritis Data Using Independent Working Correlation

The GEE Procedure

Parameter Estimates for Response Model with Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept1		-6.7528	0.4227	-7.5812	-5.9244	-15.98	<.0001
Intercept2		-4.6719	0.3953	-5.4466	-3.8972	-11.82	<.0001
Intercept3		-2.7138	0.3730	-3.4449	-1.9828	-7.28	<.0001
Intercept4		-0.4129	0.3689	-1.1360	0.3102	-1.12	0.2631
Visit	1	0.3852	0.1160	0.1578	0.6125	3.32	0.0009
Visit	3	0.3725	0.1118	0.1534	0.5916	3.33	0.0009
Visit	5	0.0000	0.0000	0.0000	0.0000	.	.
Treatment	1	0.5643	0.1679	0.2352	0.8933	3.36	0.0008
Treatment	2	0.0000	0.0000	0.0000	0.0000	.	.
Baseline	1	3.9533	0.5351	2.9046	5.0020	7.39	<.0001
Baseline	2	3.3264	0.4250	2.4934	4.1593	7.83	<.0001
Baseline	3	2.7672	0.3769	2.0285	3.5059	7.34	<.0001
Baseline	4	1.4252	0.4112	0.6192	2.2312	3.47	0.0005
Baseline	5	0.0000	0.0000	0.0000	0.0000	.	.

The QIC for the ALR model shown in Output 47.5.2 is 2241.95, whereas the QIC for the independent working correlation model shown in Output 47.5.4 is 2269.82, indicating a slightly better fit for the ALR model.

Output 47.5.4 Model Fit Criteria

GEE Fit Criteria	
QIC	2269.8166
QICu	2259.7693

Example 47.6: GEE for Nominal Multinomial Data

This example illustrates how you use the GEE procedure to analyze nominal multinomial data. A two-year study was conducted to assess the impact of access to Section 8 housing as a means of providing independent housing to the severely mentally ill homeless (Hurlbut, Wood, and Hough 1996). In this study, half of the 362 clients received Section 8 housing certificates. The assignment of Section 8 housing certificates is recorded in the variable Sec; 0 indicates clients who did not receive a certificate, and 1 indicates clients who received a certificate.

Every six months during the study, research staff interviewed all 362 clients, who provided data about their living arrangements in the previous 60 days. Clients' living arrangements were also recorded during a baseline interview. The time of interviews is recorded in the variable Time, whose value is 0, 6, 12, or 24 (for the number of months since the study began). There were a total of 159 missed interviews. The variable Housing records the living arrangement of a client and is coded as 0 (street living), 1 (community living), or 2 (independent living). The following statements create the data set Housing:

```
data Housing;
   input ID Housing Time Sec;
   datalines;
1 1 0 1
1 2 6 1
1 2 12 1
1 2 24 1
2 1 0 1
2 2 6 1

   ... more lines ...

362 1 0 0
362 1 6 0
362 1 12 0
362 1 24 0
;
```

The following SAS statements use PROC GEE to fit a model to nominal multinomial data:

```
proc gee data=Housing;
   class ID Housing Time SEC;
   model Housing=Sec / dist=multinomial link=glogit;
   repeated subject=ID / within=Time;
run;
```

An ordinary GEE that has an independent working correlation structure is fit. This model is the only option supported for data that have nominal multinomial responses. In the MODEL statement, you specify LINK=GLOGIT to indicate that the responses are nominal. In the generalized logit model, you model baseline category logits. By default, the GEE procedure chooses the last response category as the baseline category. If your nominal response has J categories, then the baseline logit for category j and subject i is

$$\log(\mu_{ij}/\mu_{iJ}) = \eta_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j$$

and

$$\mu_{ij} = \frac{\exp(\eta_{ij})}{\sum_{k=1}^{J} \exp(\eta_{ik})}$$

$$\eta_{iJ} = 0$$

The results of fitting the model are displayed in Output 47.6.1.

Output 47.6.1 Results of Model Fitting

The GEE Procedure

Parameter Estimates for Response Model with Empirical Standard Error Estimates

Parameter		Housing	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		**0**	-0.9532	0.1266	-1.2013	-0.7051	-7.53	<.0001
Intercept		**1**	-0.6562	0.1064	-0.8647	-0.4477	-6.17	<.0001
Sec	0	**0**	0.9226	0.1850	0.5599	1.2853	4.99	<.0001
Sec	0	**1**	1.2645	0.1642	0.9426	1.5863	7.70	<.0001
Sec	1	**0**	0.0000	0.0000	0.0000	0.0000	.	.
Sec	1	**1**	0.0000	0.0000	0.0000	0.0000	.	.

The positive estimates for the classification variable Sec = 0 at each response category, Housing = 0 and 1, indicate an increased probability that a client will live independently when given access to Section 8 housing. The model fit criteria are shown in Output 47.6.2

Output 47.6.2 Model Fit Criteria

GEE Fit Criteria	
QIC	2675.2174
QICu	2671.4680

For comparison, the following SAS statements treat the responses as ordinal and use PROC GEE to fit a marginal model by using an independent working correlation structure:

```
proc gee data=Housing;
   class ID Housing Time SEC;
   model Housing=Sec / dist=multinomial;
   repeated subject=ID / within=Time;
run;
```

The cumulative logit link function is the default option that is used to fit the model. Because the generalized logit link function is not specified, the responses are treated as ordinal multinomial data. The results for the model that is fit by treating the responses as ordinal are displayed in Output 47.6.3.

Output 47.6.3 Results of Model Fitting

The GEE Procedure

Parameter Estimates for Response Model with Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept1		-1.6917	0.1242	-1.9352	-1.4481	-13.62	<.0001
Intercept2		0.0112	0.0960	-0.1770	0.1994	0.12	0.9072
Sec	0	0.8224	0.1327	0.5624	1.0824	6.20	<.0001
Sec	1	0.0000	0.0000	0.0000	0.0000	.	.

Treating the responses as ordinal results in a single parameter estimate that is related to the classification variable Sec. The QIC for the model that is fit by treating the responses as nominal (shown in Output 47.6.2) is 2675.21, whereas the QIC for the model that is fit by treating the responses as ordinal (shown in Output 47.6.4) is 2710.50, indicating a slightly better fit when the responses are treated as nominal.

Output 47.6.4 Model Fit Criteria

GEE Fit Criteria	
QIC	2710.4971
QICu	2707.2983

References

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

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

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. 2nd ed. New York: Oxford University Press.

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

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995). "Regression Models for Longitudinal Binary Responses with Informative Drop-Outs." *Journal of the Royal Statistical Society, Series B* 57:691–704.

Hardin, J. W., and Hilbe, J. M. (2003). *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC.

Heagerty, P., and Zeger, S. L. (1996). "Marginal Regression Models for Clustered Ordinal Measurements." *Journal of the American Statistical Association* 91:1024–1036.

Hurlbut, M. S., Wood, P. A., and Hough, R. L. (1996). "Providing Independent Housing for the Homeless Mentally Ill: A Novel Approach to Evaluating Long-Term Longitudinal Housing Patterns." *Journal of Community Psychology* 24:291–310.

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

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.

Mallinckrodt, C. (2013). *Preventing and Treating Missing Data in Longitudinal Clinical Trials: A Practical Guide*. Cambridge: Cambridge University Press.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

Molenberghs, G., and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. New York: John Wiley & Sons.

O'Kelly, M., and Ratitch, B. (2014). *Clinical Trials with Missing Data: A Guide for Practitioners*. Chichester, UK: John Wiley & Sons.

Pan, W. (2001). "Akaike's Information Criterion in Generalized Estimating Equations." *Biometrics* 57:120–125.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. (2002). "Performance of Weighted Estimating Equations for Longitudinal Binary Data with Drop-Outs Missing at Random." *Statistics in Medicine* 21:3035–3054.

Robins, J. M., and Rotnitzky, A. (1995). "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90:122–129.

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.

Rubin, D. B. (1976). "Inference and Missing Data." *Biometrika* 63:581–592.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2012). *Categorical Data Analysis Using SAS*. 3rd ed. Cary, NC: SAS Institute Inc.

Subject Index

Syntax Index