Chapter
7
Generalized Estimating Equations

The Generalized Estimating Equations procedure extends the generalized linear model to allow for analysis of repeated measurements or other correlated observations, such as clustered data.

Example. Public health officials can use generalized estimating equations to fit a repeated measures logistic regression to study effects of air pollution on children.

Data. The response can be scale, counts, binary, or events-in-trials. Factors are assumed to be categorical. The covariates, scale weight, and offset are assumed to be scale. Variables used to define subjects or within-subject repeated measurements cannot be used to define the response but can serve other roles in the model.

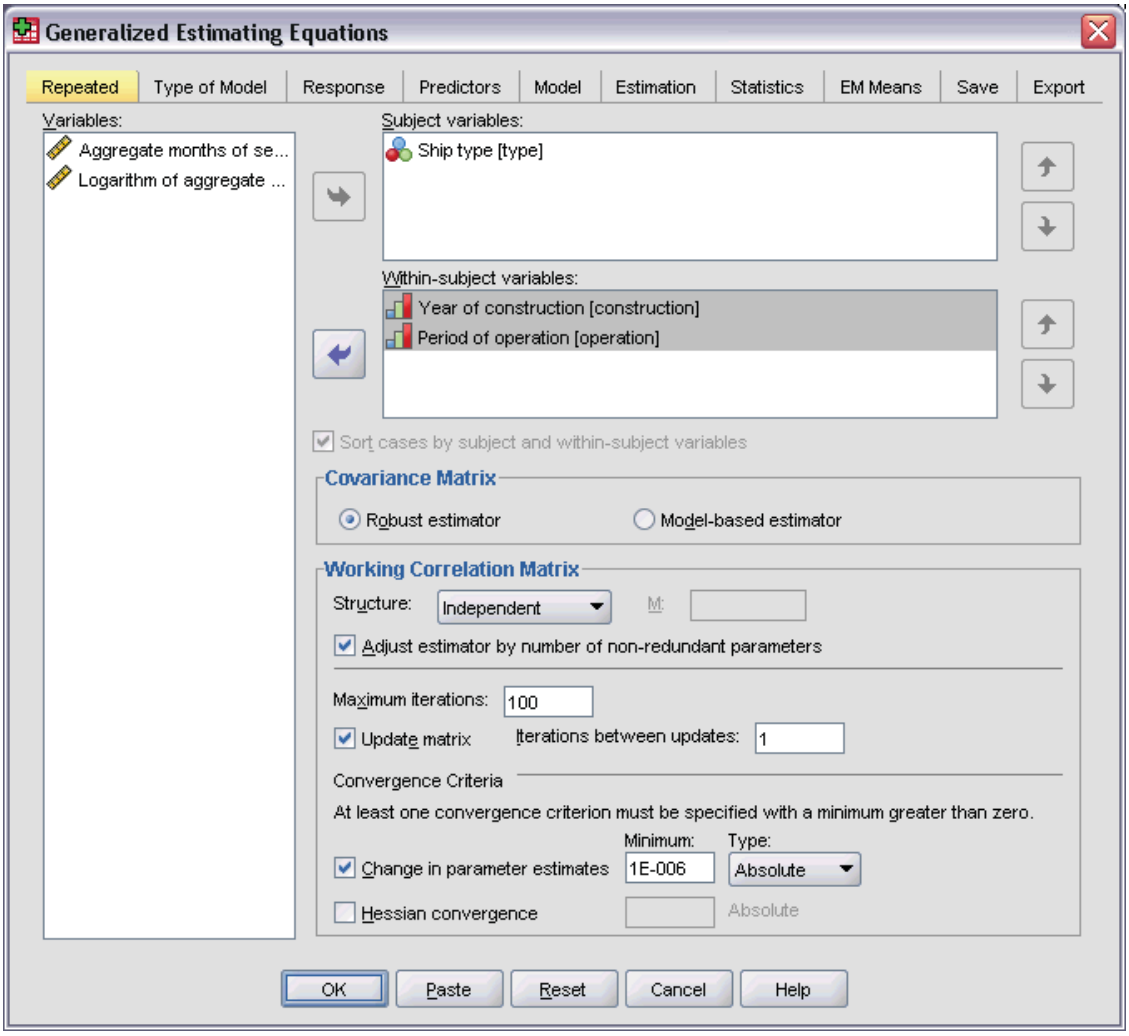
Assumptions. Cases are assumed to be dependent within subjects and independent between subjects. The correlation matrix that represents the within-subject dependencies is estimated as part of the model.

Obtaining Generalized Estimating Equations

From the menus choose:

Analyze
  Generalized Linear Models
    Generalized Estimating Equations...

Figure 7-1
Generalized Estimating Equations: Repeated tab



- Select one or more subject variables (see below for further options).

  The combination of values of the specified variables should uniquely define **subjects** within the dataset. For example, a single *Patient ID* variable should be sufficient to define subjects in a single hospital, but the combination of *Hospital ID* and *Patient ID* may be necessary if patient identification numbers are not unique across hospitals. In a repeated measures setting, multiple observations are recorded for each subject, so each subject may occupy multiple cases in the dataset.

- On the Type of Model tab, specify a distribution and link function.

- On the Response tab, select a dependent variable.
- On the Predictors tab, select factors and covariates for use in predicting the dependent variable.
- On the Model tab, specify model effects using the selected factors and covariates.

Optionally, on the Repeated tab you can specify:

Within-subject variables. The combination of values of the within-subject variables defines the ordering of measurements within subjects; thus, the combination of within-subject and subject variables uniquely defines each measurement. For example, the combination of *Period*, *Hospital ID*, and *Patient ID* defines, for each case, a particular office visit for a particular patient within a particular hospital.

If the dataset is already sorted so that each subject's repeated measurements occur in a contiguous block of cases and in the proper order, it is not strictly necessary to specify a within-subjects variable, and you can deselect Sort cases by subject and within-subject variables and save the processing time required to perform the (temporary) sort. Generally, it's a good idea to make use of within-subject variables to ensure proper ordering of measurements.

Subject and within-subject variables cannot be used to define the response, but they can perform other functions in the model. For example, *Hospital ID* could be used as a factor in the model.

Covariance Matrix. The model-based estimator is the negative of the generalized inverse of the Hessian matrix. The robust estimator (also called the Huber/White/sandwich estimator) is a "corrected" model-based estimator that provides a consistent estimate of the covariance, even when the working correlation matrix is misspecified. This specification applies to the parameters in the linear model part of the generalized estimating equations, while the specification on the Estimation tab applies only to the initial generalized linear model.

Working Correlation Matrix. This correlation matrix represents the within-subject dependencies. Its size is determined by the number of measurements and thus the combination of values of within-subject variables. You can specify one of the following structures:

- **Independent.** Repeated measurements are uncorrelated.

- **AR(1).** Repeated measurements have a first-order autoregressive relationship. The correlation between any two elements is equal to for adjacent elements, ^{2} for elements that are separated by a third, and so on. is constrained so that –1< <1.
- **Exchangeable.** This structure has homogenous correlations between elements. It is also known as a compound symmetry structure.
- **M-dependent.** Consecutive measurements have a common correlation coefficient, pairs of measurements separated by a third have a common correlation coefficient, and so on, through pairs of measurements separated by $m-1$ other measurements. Measurements with greater separation are assumed to be uncorrelated. When choosing this structure, specify a value of m less than the order of the working correlation matrix.
- **Unstructured.** This is a completely general correlation matrix.

By default, the procedure will adjust the correlation estimates by the number of nonredundant parameters. Removing this adjustment may be desirable if you want the estimates to be invariant to subject-level replication changes in the data.

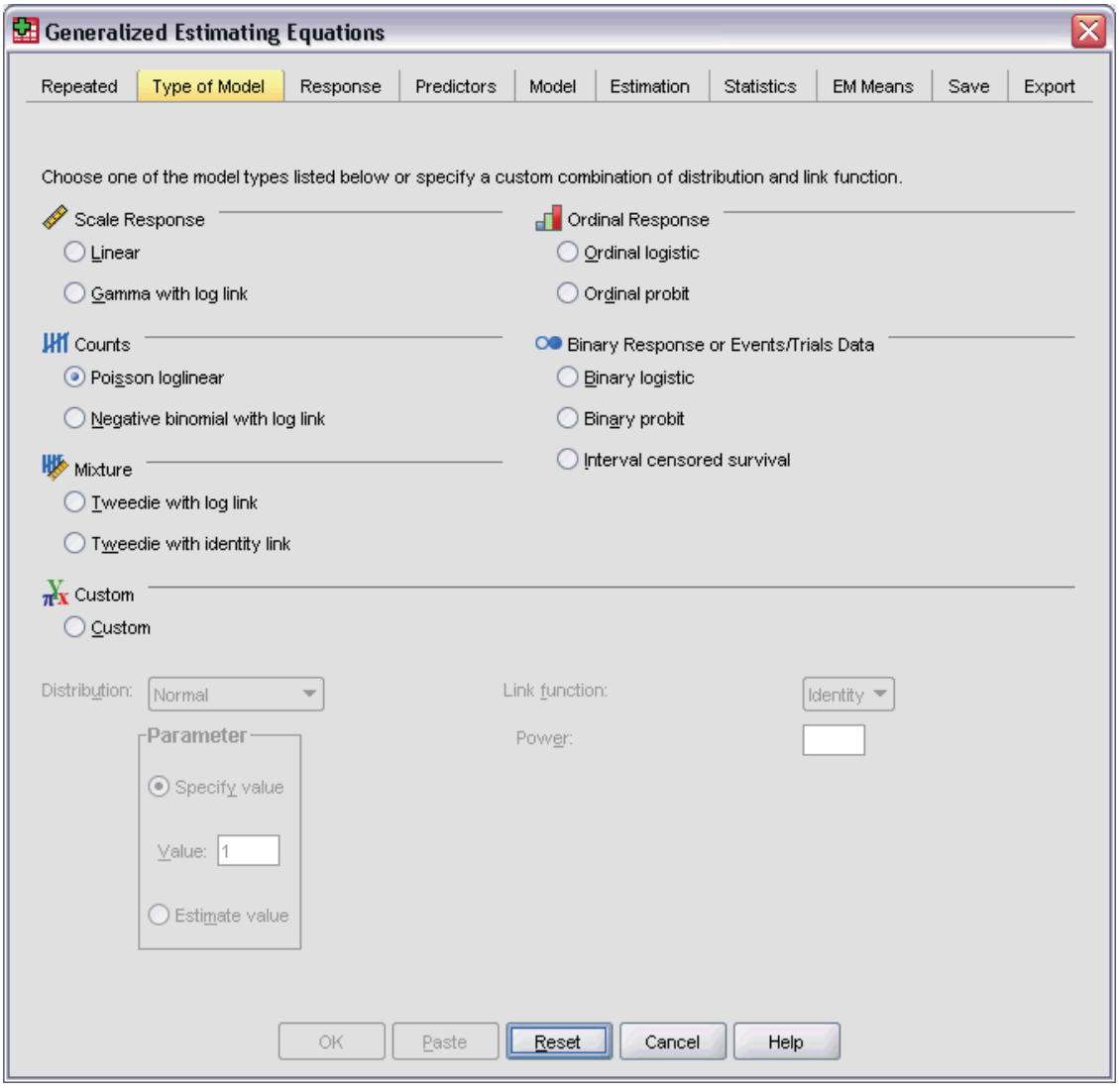
- **Maximum iterations.** The maximum number of iterations the generalized estimating equations algorithm will execute. Specify a non-negative integer. This specification applies to the parameters in the linear model part of the generalized estimating equations, while the specification on the Estimation tab applies only to the initial generalized linear model.
- **Update matrix.** Elements in the working correlation matrix are estimated based on the parameter estimates, which are updated in each iteration of the algorithm. If the working correlation matrix is not updated at all, the initial working correlation matrix is used throughout the estimation process. If the matrix is updated, you can specify the iteration interval at which to update working correlation matrix elements. Specifying a value greater than 1 may reduce processing time.

Convergence criteria. These specifications apply to the parameters in the linear model part of the generalized estimating equations, while the specification on the Estimation tab applies only to the initial generalized linear model.

- **Parameter convergence.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the parameter estimates is less than the value specified, which must be positive.
- **Hessian convergence.** Convergence is assumed if a statistic based on the Hessian is less than the value specified, which must be positive.

Generalized Estimating Equations Type of Model

Figure 7-2
Generalized Estimating Equations: Type of Model tab



The Type of Model tab allows you to specify the distribution and link function for your model, providing shortcuts for several common models that are categorized by response type.

Model Types

Scale Response.

- **Linear.** Specifies Normal as the distribution and Identity as the link function.
- **Gamma with log link.** Specifies Gamma as the distribution and Log as the link function.

Ordinal Response.

- **Ordinal logistic.** Specifies Multinomial (ordinal) as the distribution and Cumulative logit as the link function.
- **Ordinal probit.** Specifies Multinomial (ordinal) as the distribution and Cumulative probit as the link function.

Counts.

- **Poisson loglinear.** Specifies Poisson as the distribution and Log as the link function.
- **Negative binomial with log link.** Specifies Negative binomial (with a value of 1 for the ancillary parameter) as the distribution and Log as the link function. To have the procedure estimate the value of the ancillary parameter, specify a custom model with Negative binomial distribution and select Estimate value in the Parameter group.

Binary Response or Events/Trials Data.

- **Binary logistic.** Specifies Binomial as the distribution and Logit as the link function.
- **Binary probit.** Specifies Binomial as the distribution and Probit as the link function.
- **Interval censored survival.** Specifies Binomial as the distribution and Complementary log-log as the link function.

Mixture.

- **Tweedie with log link.** Specifies Tweedie as the distribution and Log as the link function.
- **Tweedie with identity link.** Specifies Tweedie as the distribution and Identity as the link function.

Custom. Specify your own combination of distribution and link function.

Distribution

This selection specifies the distribution of the dependent variable. The ability to specify a non-normal distribution and non-identity link function is the essential improvement of the generalized linear model over the general linear model. There are many possible distribution-link function combinations, and several may be appropriate for any given dataset, so your choice can be guided by a priori theoretical considerations or which combination seems to fit best.

- **Binomial.** This distribution is appropriate only for variables that represent a binary response or number of events.
- **Gamma.** This distribution is appropriate for variables with positive scale values that are skewed toward larger positive values. If a data value is less than or equal to 0 or is missing, then the corresponding case is not used in the analysis.
- **Inverse Gaussian.** This distribution is appropriate for variables with positive scale values that are skewed toward larger positive values. If a data value is less than or equal to 0 or is missing, then the corresponding case is not used in the analysis.
- **Multinomial.** This distribution is appropriate for variables that represent an ordinal response. The dependent variable can be numeric or string, and it must have at least two distinct valid data values.
- **Negative binomial.** This distribution can be thought of as the number of trials required to observe *k* successes and is appropriate for variables with non-negative integer values. If a data value is non-integer, less than 0, or missing, then the corresponding case is not used in the analysis. The value of the negative binomial distribution's ancillary parameter can be any number greater than or equal to 0; you can set it to a fixed value or allow it to be estimated by the procedure. When the ancillary parameter is set to 0, using this distribution is equivalent to using the Poisson distribution.
- **Normal.** This is appropriate for scale variables whose values take a symmetric, bell-shaped distribution about a central (mean) value. The dependent variable must be numeric.

- **Poisson.** This distribution can be thought of as the number of occurrences of an event of interest in a fixed period of time and is appropriate for variables with non-negative integer values. If a data value is non-integer, less than 0, or missing, then the corresponding case is not used in the analysis.
- **Tweedie.** This distribution is appropriate for variables that can be represented by Poisson mixtures of gamma distributions; the distribution is "mixed" in the sense that it combines properties of continuous (takes non-negative real values) and discrete distributions (positive probability mass at a single value, 0). The dependent variable must be numeric, with data values greater than or equal to zero. If a data value is less than zero or missing, then the corresponding case is not used in the analysis. The fixed value of the Tweedie distribution's parameter can be any number greater than one and less than two.

Link Function

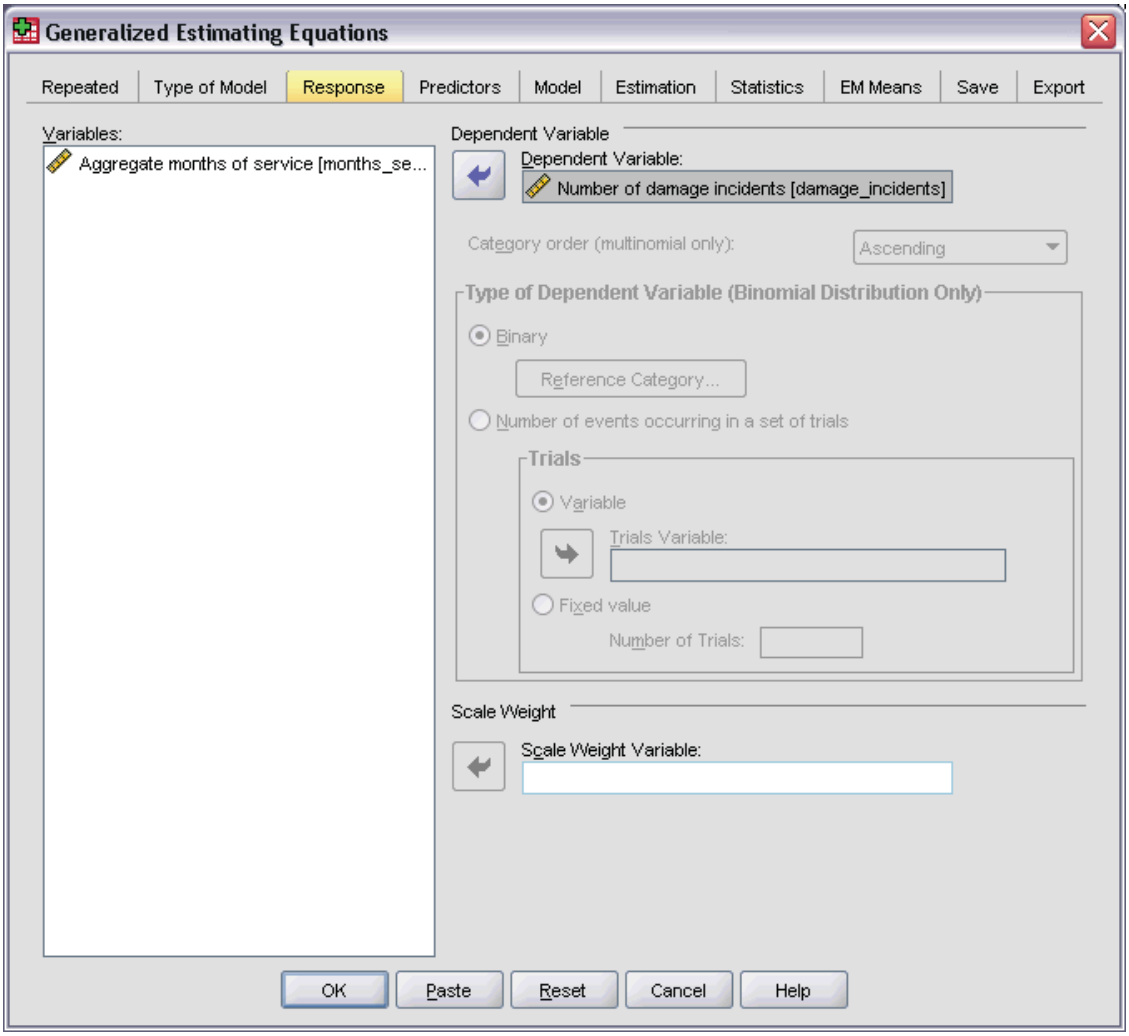
The link function is a transformation of the dependent variable that allows estimation of the model. The following functions are available:

- **Identity.** $f(x)=x$. The dependent variable is not transformed. This link can be used with any distribution.
- **Complementary log-log.** $f(x)=\log(-\log(1-x))$. This is appropriate only with the binomial distribution.
- **Cumulative Cauchit.** $f(x) = \tan(\pi\ (x - 0.5))$, applied to the cumulative probability of each category of the response. This is appropriate only with the multinomial distribution.
- **Cumulative complementary log-log.** $f(x)=\ln(-\ln(1-x))$, applied to the cumulative probability of each category of the response. This is appropriate only with the multinomial distribution.
- **Cumulative logit.** $f(x)=\ln(x\ /\ (1-x))$, applied to the cumulative probability of each category of the response. This is appropriate only with the multinomial distribution.
- **Cumulative negative log-log.** $f(x)=-\ln(-\ln(x))$, applied to the cumulative probability of each category of the response. This is appropriate only with the multinomial distribution.
- **Cumulative probit.** $f(x)=\Phi^{-1}(x)$, applied to the cumulative probability of each category of the response, where Φ^{-1} is the inverse standard normal cumulative distribution function. This is appropriate only with the multinomial distribution.

- **Log.** $f(x)=\log(x)$. This link can be used with any distribution.
- **Log complement.** $f(x)=\log(1-x)$. This is appropriate only with the binomial distribution.
- **Logit.** $f(x)=\log(x / (1-x))$. This is appropriate only with the binomial distribution.
- **Negative binomial.** $f(x)=\log(x / (x+k^{-1}))$, where k is the ancillary parameter of the negative binomial distribution. This is appropriate only with the negative binomial distribution.
- **Negative log-log.** $f(x)=-\log(-\log(x))$. This is appropriate only with the binomial distribution.
- **Odds power.** $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$, if $\alpha \neq 0$. $f(x)=\log(x)$, if $\alpha=0$. α is the required number specification and must be a real number. This is appropriate only with the binomial distribution.
- **Probit.** $f(x)=\Phi^{-1}(x)$, where Φ^{-1} is the inverse standard normal cumulative distribution function. This is appropriate only with the binomial distribution.
- **Power.** $f(x)=x^{\alpha}$, if $\alpha \neq 0$. $f(x)=\log(x)$, if $\alpha=0$. α is the required number specification and must be a real number. This link can be used with any distribution.

Generalized Estimating Equations Response

Figure 7-3
Generalized Estimating Equations: Response tab



In many cases, you can simply specify a dependent variable; however, variables that take only two values and responses that record events in trials require extra attention.

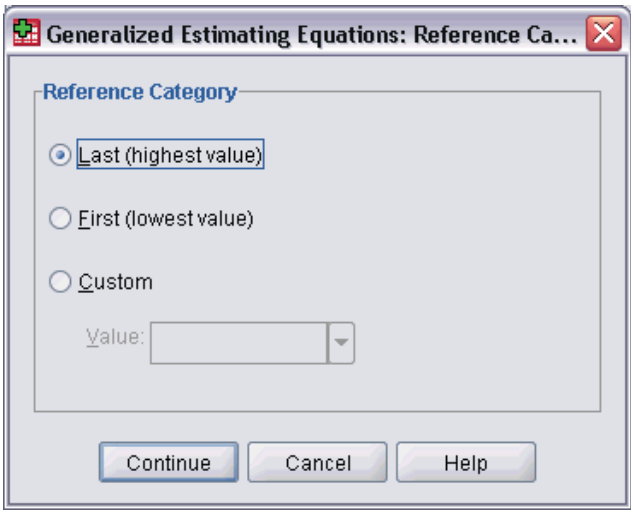
- **Binary response.** When the dependent variable takes only two values, you can specify the reference category for parameter estimation. A binary response variable can be string or numeric.
- **Number of events occurring in a set of trials.** When the response is a number of events occurring in a set of trials, the dependent variable contains the number of events and you can select an additional variable containing the number of trials. Alternatively, if the number of trials is the same across all subjects, then trials may be specified using a fixed value. The number of trials should be greater than or equal to the number of events for each case. Events should be non-negative integers, and trials should be positive integers.

For ordinal multinomial models, you can specify the category order of the response: ascending, descending, or data (data order means that the first value encountered in the data defines the first category, the last value encountered defines the last category).

Scale Weight. The scale parameter is an estimated model parameter related to the variance of the response. The scale weights are "known" values that can vary from observation to observation. If the scale weight variable is specified, the scale parameter, which is related to the variance of the response, is divided by it for each observation. Cases with scale weight values that are less than or equal to 0 or are missing are not used in the analysis.

Generalized Estimating Equations Reference Category

Figure 7-4
Generalized Estimating Equations Reference Category dialog box

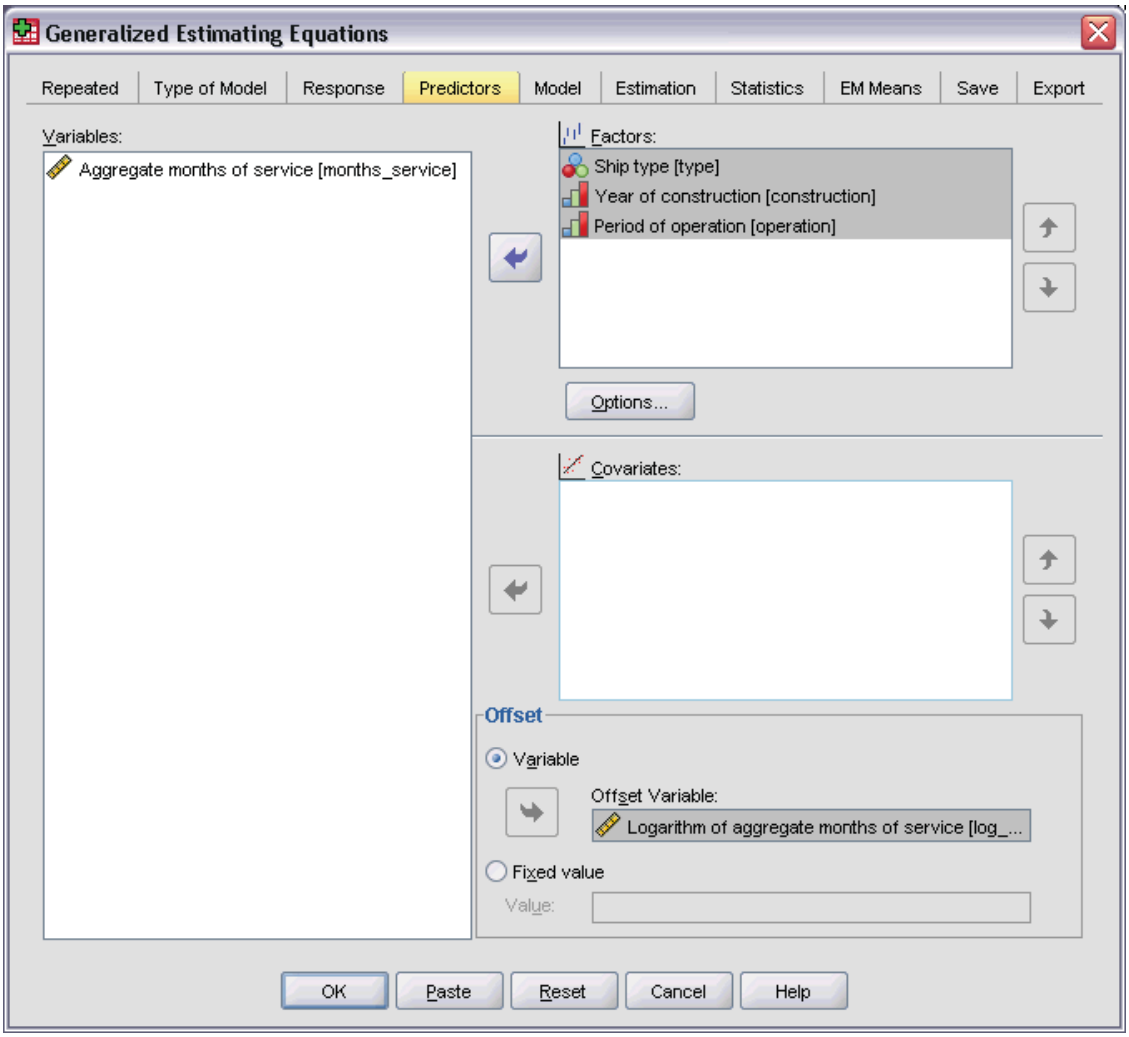


For binary response, you can choose the reference category for the dependent variable. This can affect certain output, such as parameter estimates and saved values, but it should not change the model fit. For example, if your binary response takes values 0 and 1:

- By default, the procedure makes the last (highest-valued) category, or 1, the reference category. In this situation, model-saved probabilities estimate the chance that a given case takes the value 0, and parameter estimates should be interpreted as relating to the likelihood of category 0.
- If you specify the first (lowest-valued) category, or 0, as the reference category, then model-saved probabilities estimate the chance that a given case takes the value 1.
- If you specify the custom category and your variable has defined labels, you can set the reference category by choosing a value from the list. This can be convenient when, in the middle of specifying a model, you don't remember exactly how a particular variable was coded.

Generalized Estimating Equations Predictors

Figure 7-5
Generalized Estimating Equations: Predictors tab



The Predictors tab allows you to specify the factors and covariates used to build model effects and to specify an optional offset.

Factors. Factors are categorical predictors; they can be numeric or string.

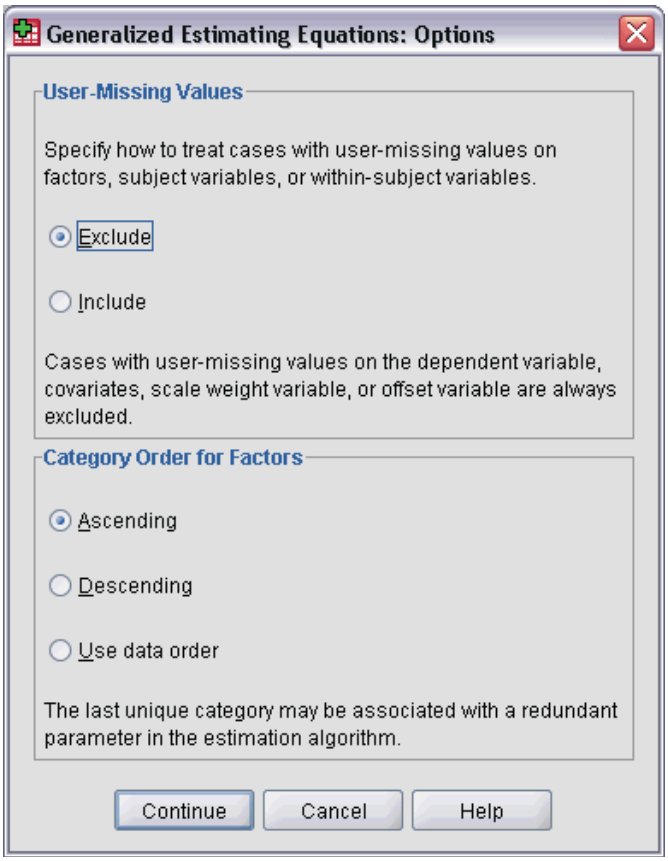
Covariates. Covariates are scale predictors; they must be numeric.

Note: When the response is binomial with binary format, the procedure computes deviance and chi-square goodness-of-fit statistics by subpopulations that are based on the cross-classification of observed values of the selected factors and covariates. You should keep the same set of predictors across multiple runs of the procedure to ensure a consistent number of subpopulations.

Offset. The offset term is a "structural" predictor. Its coefficient is not estimated by the model but is assumed to have the value 1; thus, the values of the offset are simply added to the linear predictor of the dependent variable. This is especially useful in Poisson regression models, where each case may have different levels of exposure to the event of interest. For example, when modeling accident rates for individual drivers, there is an important difference between a driver who has been at fault in one accident in three years of experience and a driver who has been at fault in one accident in 25 years! The number of accidents can be modeled as a Poisson response if the experience of the driver is included as an offset term.

Generalized Estimating Equations Options

Figure 7-6
Generalized Estimating Equations Options dialog box



These options are applied to all factors specified on the Predictors tab.

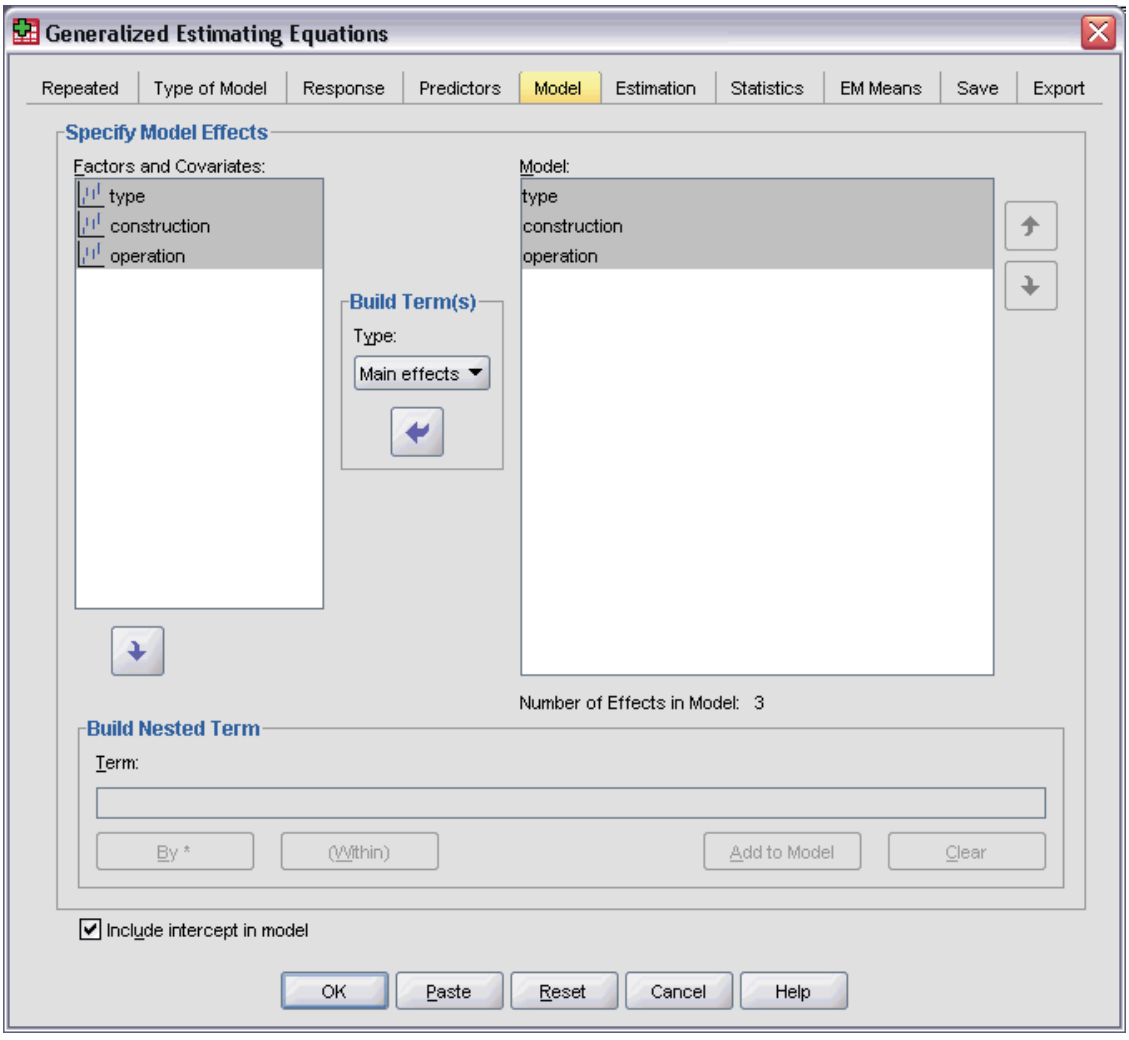
User-Missing Values. Factors must have valid values for a case to be included in the analysis. These controls allow you to decide whether user-missing values are treated as valid among factor variables.

Category Order. This is relevant for determining a factor's last level, which may be associated with a redundant parameter in the estimation algorithm. Changing the category order can change the values of factor-level effects, since these parameter estimates are calculated relative to the "last" level. Factors can be sorted in ascending order from lowest to highest value, in descending order from highest to lowest value,

or in "data order." This means that the first value encountered in the data defines the first category, and the last unique value encountered defines the last category.

Generalized Estimating Equations Model

Figure 7-7
Generalized Estimating Equations: Model tab



Specify Model Effects. The default model is intercept-only, so you must explicitly specify other model effects. Alternatively, you can build nested or non-nested terms.

Non-Nested Terms

For the selected factors and covariates:

Main effects. Creates a main-effects term for each variable selected.

Interaction. Creates the highest-level interaction term for all selected variables.

Factorial. Creates all possible interactions and main effects of the selected variables.

All 2-way. Creates all possible two-way interactions of the selected variables.

All 3-way. Creates all possible three-way interactions of the selected variables.

All 4-way. Creates all possible four-way interactions of the selected variables.

All 5-way. Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *Customer* effect can be said to be **nested within** the *Store location* effect.

Additionally, you can include interaction effects or add multiple levels of nesting to the nested term.

Limitations. Nested terms have the following restrictions:

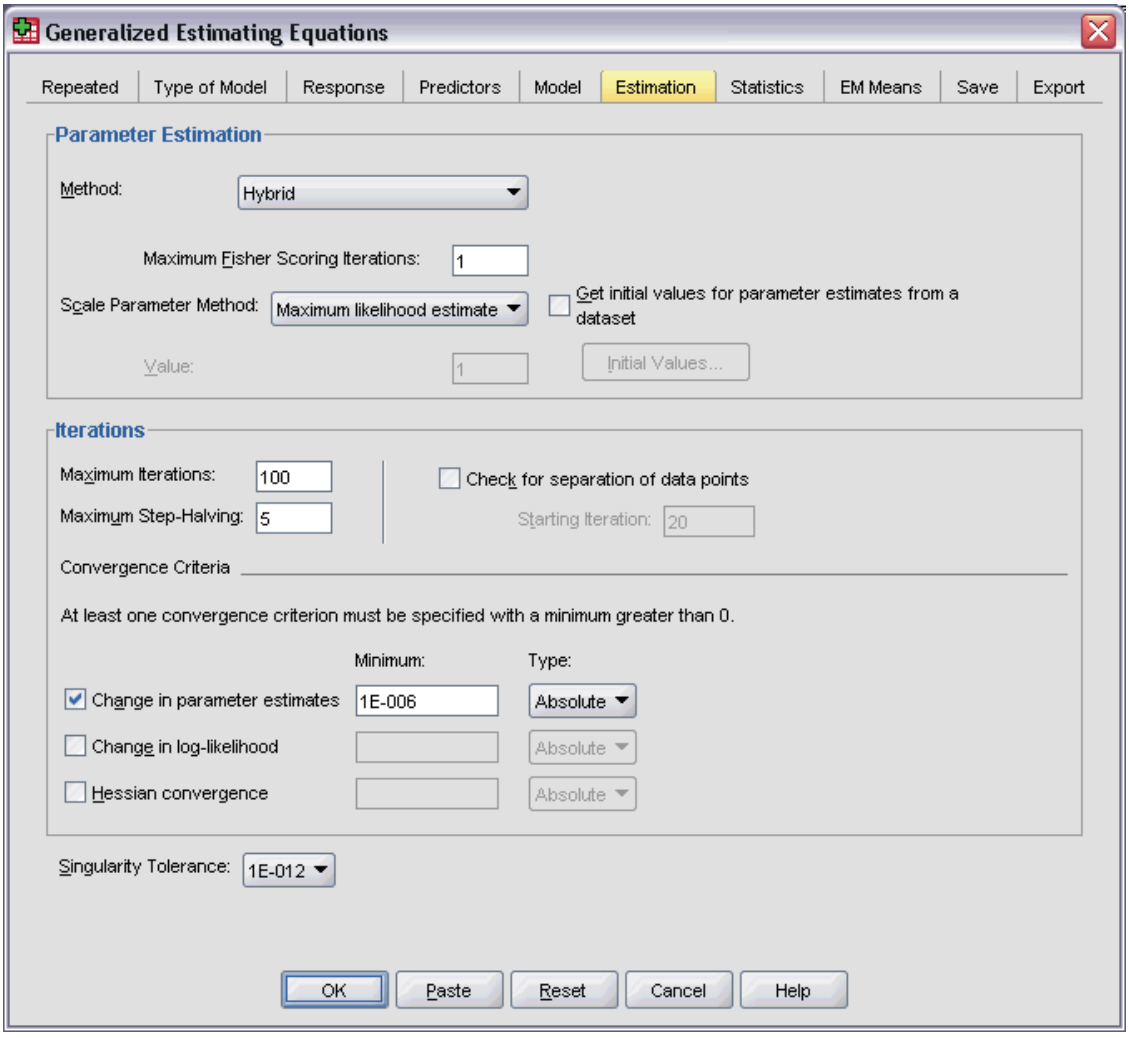
- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A*A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

Intercept. The intercept is usually included in the model. If you can assume the data pass through the origin, you can exclude the intercept.

Models with the multinomial ordinal distribution do not have a single intercept term; instead there are threshold parameters that define transition points between adjacent categories. The thresholds are always included in the model.

Generalized Estimating Equations Estimation

Figure 7-8
Generalized Estimating Equations: Estimation tab



Parameter Estimation. The controls in this group allow you to specify estimation methods and to provide initial values for the parameter estimates.

- **Method.** You can select a parameter estimation method; choose between Newton-Raphson, Fisher scoring, or a hybrid method in which Fisher scoring iterations are performed before switching to the Newton-Raphson method. If convergence is achieved during the Fisher scoring phase of the hybrid method

before the maximum number of Fisher iterations is reached, the algorithm continues with the Newton-Raphson method.

- **Scale Parameter Method.** You can select the scale parameter estimation method.

  Maximum-likelihood jointly estimates the scale parameter with the model effects; note that this option is not valid if the response has a negative binomial, Poisson, or binomial distribution. Since the concept of likelihood does not enter into generalized estimating equations, this specification applies only to the initial generalized linear model; this scale parameter estimate is then passed to the generalized estimating equations, which update the scale parameter by the Pearson chi-square divided by its degrees of freedom.

  The deviance and Pearson chi-square options estimate the scale parameter from the value of those statistics in the initial generalized linear model; this scale parameter estimate is then passed to the generalized estimating equations, which treat it as fixed.

  Alternatively, specify a fixed value for the scale parameter. It will be treated as fixed in estimating the initial generalized linear model and the generalized estimating equations.

- **Initial values.** The procedure will automatically compute initial values for parameters. Alternatively, you can specify initial values for the parameter estimates.

The iterations and convergence criteria specified on this tab are applicable only to the initial generalized linear model. For estimation criteria used in fitting the generalized estimating equations, see the Repeated tab.

Iterations.

- **Maximum iterations.** The maximum number of iterations the algorithm will execute. Specify a non-negative integer.
- **Maximum step-halving.** At each iteration, the step size is reduced by a factor of 0.5 until the log-likelihood increases or maximum step-halving is reached. Specify a positive integer.
- **Check for separation of data points.** When selected, the algorithm performs tests to ensure that the parameter estimates have unique values. Separation occurs when the procedure can produce a model that correctly classifies every case. This option is available for multinomial responses and binomial responses with binary format.

Convergence Criteria.

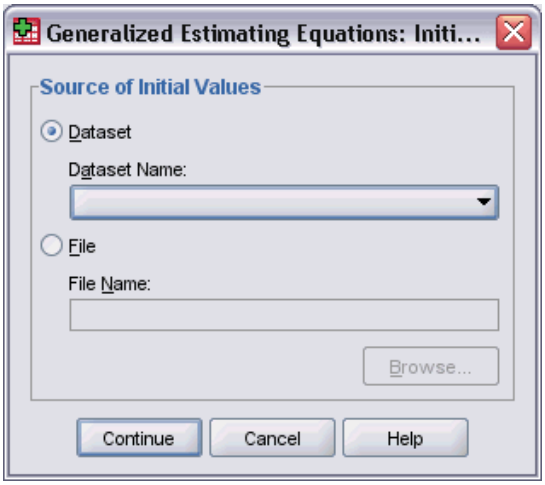
- **Parameter convergence.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the parameter estimates is less than the value specified, which must be positive.
- **Log-likelihood convergence.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the log-likelihood function is less than the value specified, which must be positive.
- **Hessian convergence.** For the Absolute specification, convergence is assumed if a statistic based on the Hessian convergence is less than the positive value specified. For the Relative specification, convergence is assumed if the statistic is less than the product of the positive value specified and the absolute value of the log-likelihood.

Singularity tolerance. Singular (or non-invertible) matrices have linearly dependent columns, which can cause serious problems for the estimation algorithm. Even near-singular matrices can lead to poor results, so the procedure will treat a matrix whose determinant is less than the tolerance as singular. Specify a positive value.

Generalized Estimating Equations Initial Values

The procedure estimates an initial generalized linear model, and the estimates from this model are used as initial values for the parameter estimates in the linear model part of the generalized estimating equations. Initial values are not needed for the working correlation matrix because matrix elements are based on the parameter estimates. Initial values specified on this dialog box are used as the starting point for the initial generalized linear model, not the generalized estimating equations, unless the Maximum iterations on the Estimation tab is set to 0.

Figure 7-9
Generalized Estimating Equations Initial Values dialog box



If initial values are specified, they must be supplied for all parameters (including redundant parameters) in the model. In the dataset, the ordering of variables from left to right must be: *RowType_*, *VarName_*, *P1*, *P2*, …, where *RowType_* and *VarName_* are string variables and *P1*, *P2*, … are numeric variables corresponding to an ordered list of the parameters.

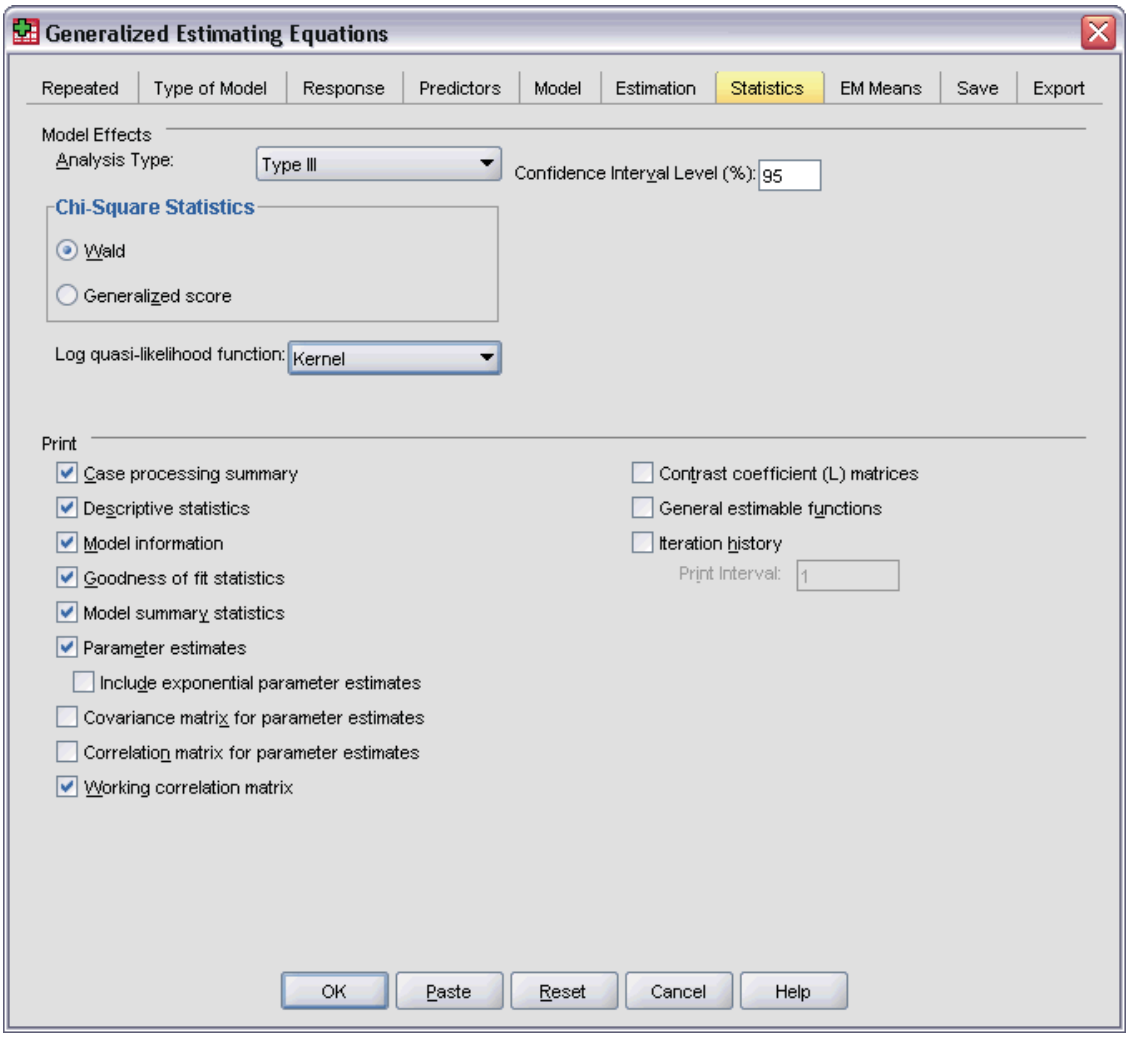
- Initial values are supplied on a record with value *EST* for variable *RowType_*; the actual initial values are given under variables *P1*, *P2*, …. The procedure ignores all records for which *RowType_* has a value other than *EST* as well as any records beyond the first occurrence of *RowType_* equal to *EST*.
- The intercept, if included in the model, or threshold parameters, if the response has a multinomial distribution, must be the first initial values listed.
- The scale parameter and, if the response has a negative binomial distribution, the negative binomial parameter, must be the last initial values specified.
- If Split File is in effect, then the variables must begin with the split-file variable or variables in the order specified when creating the Split File, followed by *RowType_*, *VarName_*, *P1*, *P2*, … as above. Splits must occur in the specified dataset in the same order as in the original dataset.

Note: The variable names *P1*, *P2*, … are not required; the procedure will accept any valid variable names for the parameters because the mapping of variables to parameters is based on variable position, not variable name. Any variables beyond the last parameter are ignored.

The file structure for the initial values is the same as that used when exporting the model as data; thus, you can use the final values from one run of the procedure as input in a subsequent run.

Generalized Estimating Equations Statistics

Figure 7-10
Generalized Estimating Equations: Statistics tab



Model Effects.

- **Analysis type.** Specify the type of analysis to produce for testing model effects. Type I analysis is generally appropriate when you have a priori reasons for ordering predictors in the model, while Type III is more generally applicable. Wald or generalized score statistics are computed based upon the selection in the Chi-Square Statistics group.
- **Confidence intervals.** Specify a confidence level greater than 50 and less than 100. Wald intervals are always produced regardless of the type of chi-square statistics selected, and are based on the assumption that parameters have an asymptotic normal distribution.
- **Log quasi-likelihood function.** This controls the display format of the log quasi-likelihood function. The full function includes an additional term that is constant with respect to the parameter estimates; it has no effect on parameter estimation and is left out of the display in some software products.

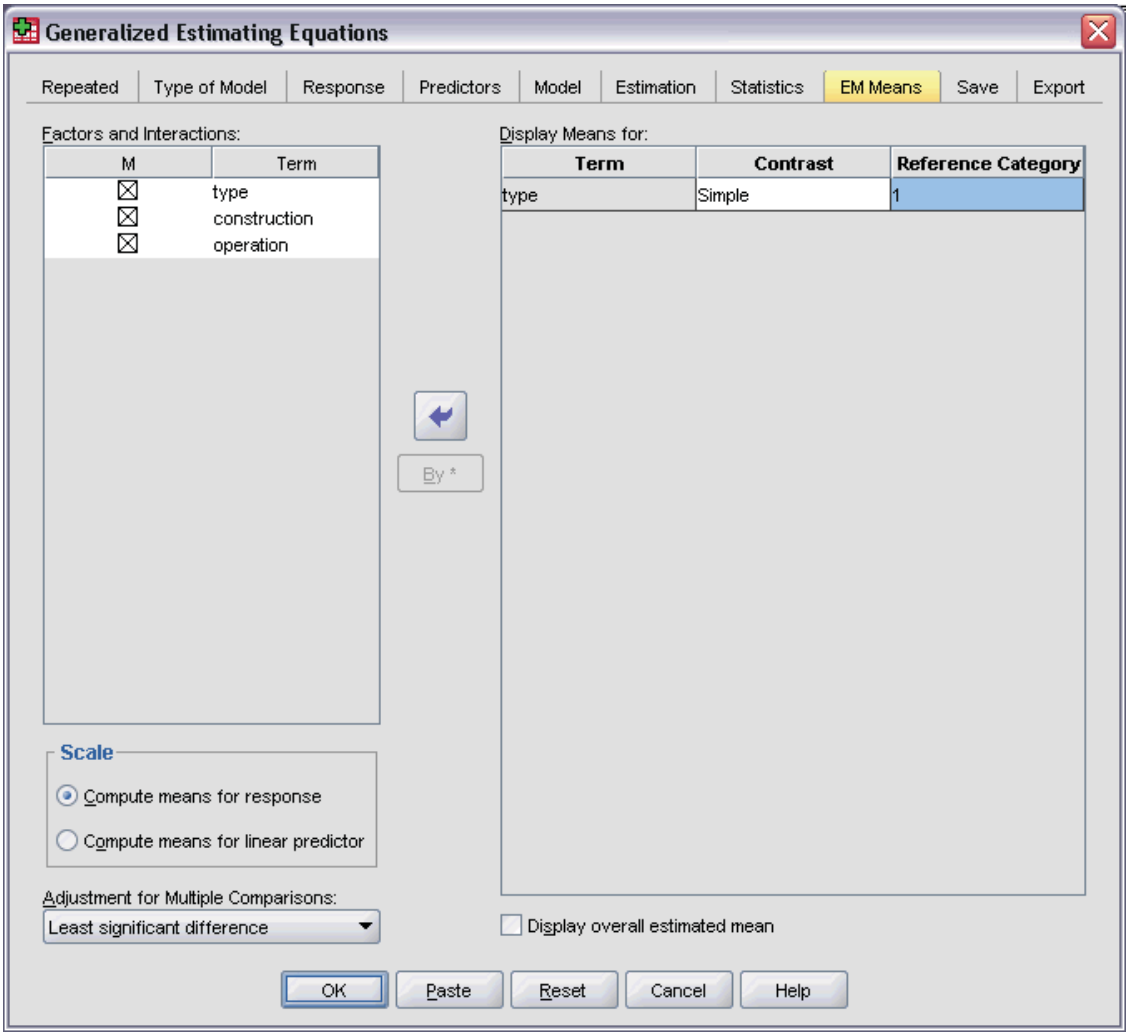
Print. The following output is available.

- **Case processing summary.** Displays the number and percentage of cases included and excluded from the analysis and the Correlated Data Summary table.
- **Descriptive statistics.** Displays descriptive statistics and summary information about the dependent variable, covariates, and factors.
- **Model information.** Displays the dataset name, dependent variable or events and trials variables, offset variable, scale weight variable, probability distribution, and link function.
- **Goodness of fit statistics.** Displays two extensions of Akaike's Information Criterion for model selection: Quasi-likelihood under the independence model criterion (QIC) for choosing the best correlation structure and another QIC measure for choosing the best subset of predictors.
- **Model summary statistics.** Displays model fit tests, including likelihood-ratio statistics for the model fit omnibus test and statistics for the Type I or III contrasts for each effect.
- **Parameter estimates.** Displays parameter estimates and corresponding test statistics and confidence intervals. You can optionally display exponentiated parameter estimates in addition to the raw parameter estimates.
- **Covariance matrix for parameter estimates.** Displays the estimated parameter covariance matrix.
- **Correlation matrix for parameter estimates.** Displays the estimated parameter correlation matrix.

- **Contrast coefficient (L) matrices.** Displays contrast coefficients for the default effects and for the estimated marginal means, if requested on the EM Means tab.
- **General estimable functions.** Displays the matrices for generating the contrast coefficient (L) matrices.
- **Iteration history.** Displays the iteration history for the parameter estimates and log-likelihood and prints the last evaluation of the gradient vector and the Hessian matrix. The iteration history table displays parameter estimates for every n^{th} iterations beginning with the 0^{th} iteration (the initial estimates), where *n* is the value of the print interval. If the iteration history is requested, then the last iteration is always displayed regardless of *n*.
- **Working correlation matrix.** Displays the values of the matrix representing the within-subject dependencies. Its structure depends upon the specifications in the Repeated tab.

Generalized Estimating Equations EM Means

Figure 7-11
Generalized Estimating Equations: EM Means tab



This tab allows you to display the estimated marginal means for levels of factors and factor interactions. You can also request that the overall estimated mean be displayed. Estimated marginal means are not available for ordinal multinomial models.

Factors and Interactions. This list contains factors specified on the Predictors tab and factor interactions specified on the Model tab. Covariates are excluded from this list. Terms can be selected directly from this list or combined into an interaction term using the By * button.

Display Means For. Estimated means are computed for the selected factors and factor interactions. The contrast determines how hypothesis tests are set up to compare the estimated means. The simple contrast requires a reference category or factor level against which the others are compared.

- **Pairwise.** Pairwise comparisons are computed for all-level combinations of the specified or implied factors. This is the only available contrast for factor interactions.
- **Simple.** Compares the mean of each level to the mean of a specified level. This type of contrast is useful when there is a control group.
- **Deviation.** Each level of the factor is compared to the grand mean. Deviation contrasts are not orthogonal.
- **Difference.** Compares the mean of each level (except the first) to the mean of previous levels. They are sometimes called reverse Helmert contrasts.
- **Helmert.** Compares the mean of each level of the factor (except the last) to the mean of subsequent levels.
- **Repeated.** Compares the mean of each level (except the last) to the mean of the subsequent level.
- **Polynomial.** Compares the linear effect, quadratic effect, cubic effect, and so on. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; and so on. These contrasts are often used to estimate polynomial trends.

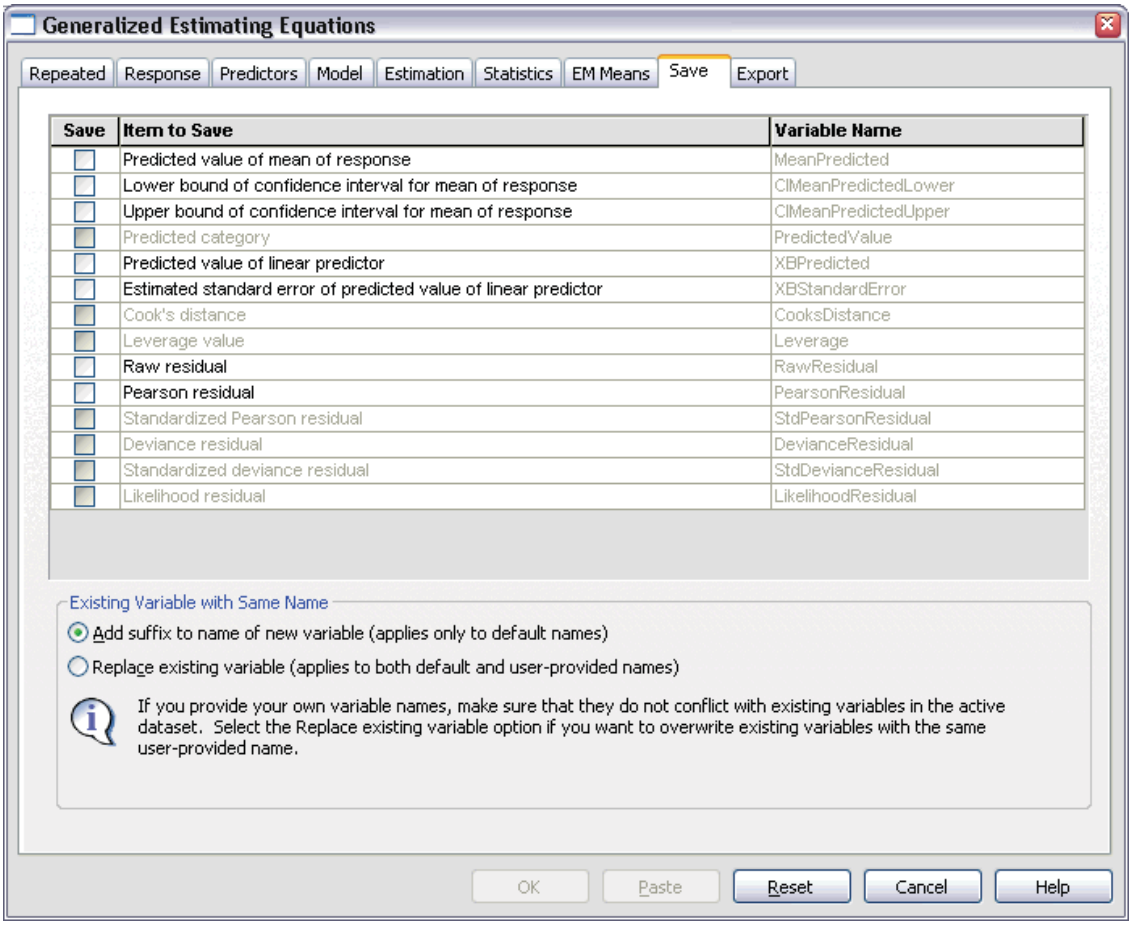
Scale. Estimated marginal means can be computed for the response, based on the original scale of the dependent variable, or for the linear predictor, based on the dependent variable as transformed by the link function.

Adjustment for Multiple Comparisons. When performing hypothesis tests with multiple contrasts, the overall significance level can be adjusted from the significance levels for the included contrasts. This group allows you to choose the adjustment method.

- **Least significant difference.** This method does not control the overall probability of rejecting the hypotheses that some linear contrasts are different from the null hypothesis values.
- **Bonferroni.** This method adjusts the observed significance level for the fact that multiple contrasts are being tested.
- **Sequential Bonferroni.** This is a sequentially step-down rejective Bonferroni procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- **Sidak.** This method provides tighter bounds than the Bonferroni approach.
- **Sequential Sidak.** This is a sequentially step-down rejective Sidak procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.

Generalized Estimating Equations Save

Figure 7-12
Generalized Estimating Equations: Save tab



Checked items are saved with the specified name; you can choose to overwrite existing variables with the same name as the new variables or avoid name conflicts by appendix suffixes to make the new variable names unique.

- **Predicted value of mean of response.** Saves model-predicted values for each case in the original response metric. When the response distribution is binomial and the dependent variable is binary, the procedure saves predicted probabilities. When the response distribution is multinomial, the item label becomes Cumulative predicted probability, and the procedure saves the cumulative predicted probability for each category of the response, except the last, up to the number of specified categories to save.

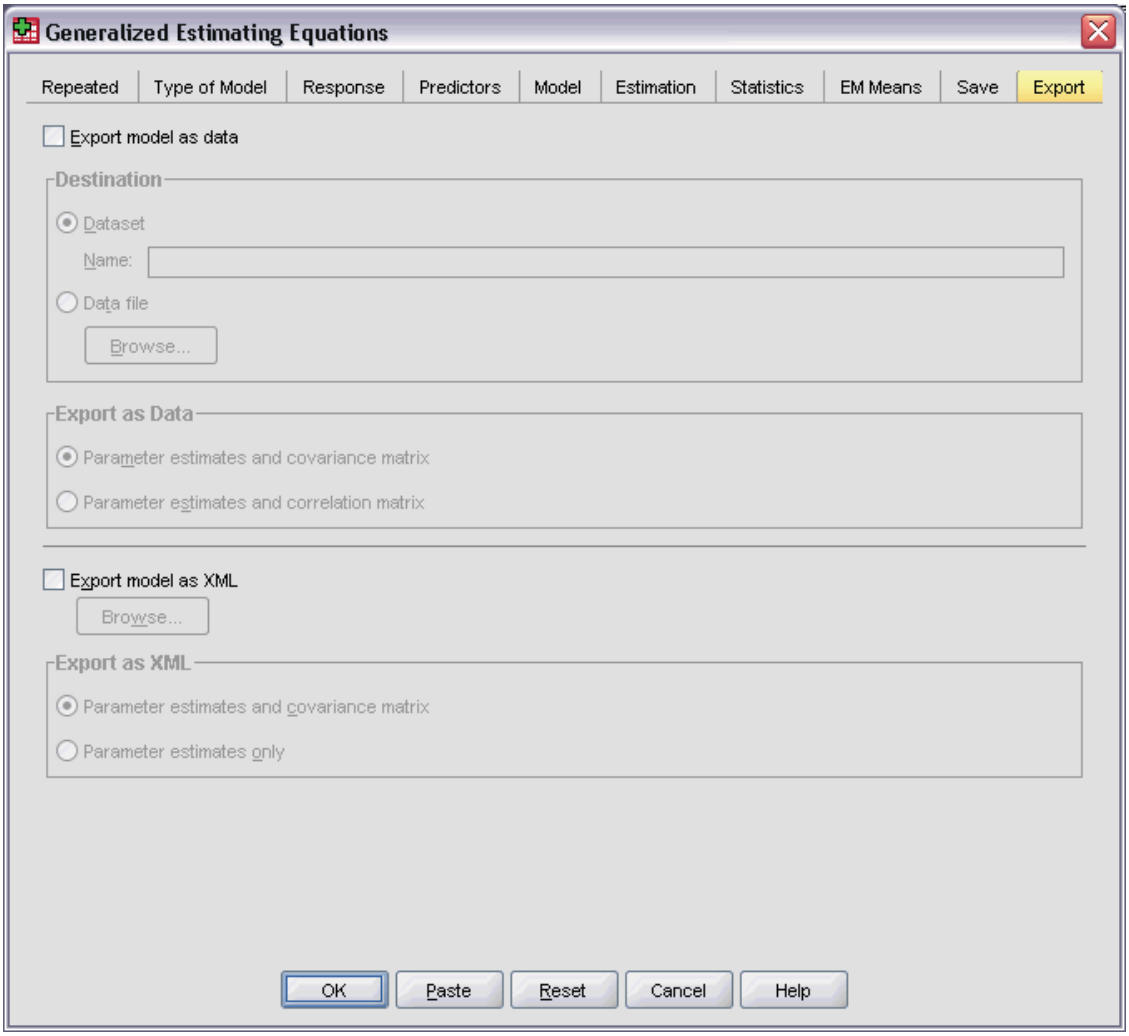
- **Lower bound of confidence interval for mean of response.** Saves the lower bound of the confidence interval for the mean of the response. When the response distribution is multinomial, the item label becomes Lower bound of confidence interval for cumulative predicted probability, and the procedure saves the lower bound for each category of the response, except the last, up to the number of specified categories to save.
- **Upper bound of confidence interval for mean of response.** Saves the upper bound of the confidence interval for the mean of the response. When the response distribution is multinomial, the item label becomes Upper bound of confidence interval for cumulative predicted probability, and the procedure saves the upper bound for each category of the response, except the last, up to the number of specified categories to save.
- **Predicted category.** For models with binomial distribution and binary dependent variable, or multinomial distribution, this saves the predicted response category for each case. This option is not available for other response distributions.
- **Predicted value of linear predictor.** Saves model-predicted values for each case in the metric of the linear predictor (transformed response via the specified link function). When the response distribution is multinomial, the procedure saves the predicted value for each category of the response, except the last, up to the number of specified categories to save.
- **Estimated standard error of predicted value of linear predictor.** When the response distribution is multinomial, the procedure saves the estimated standard error for each category of the response, except the last, up to the number of specified categories to save.

The following items are not available when the response distribution is multinomial.

- **Raw residual.** The difference between an observed value and the value predicted by the model.
- **Pearson residual.** The square root of the contribution of a case to the Pearson chi-square statistic, with the sign of the raw residual.

Generalized Estimating Equations Export

Figure 7-13
Generalized Estimating Equations: Export tab



Export model as data. Writes an SPSS Statistics dataset containing the parameter correlation or covariance matrix with parameter estimates, standard errors, significance values, and degrees of freedom. The order of variables in the matrix file is as follows.

- **Split variables.** If used, any variables defining splits.
- **RowType_.** Takes values (and value labels) *COV* (covariances), *CORR* (correlations), *EST* (parameter estimates), *SE* (standard errors), *SIG* (significance levels), and *DF* (sampling design degrees of freedom). There is a separate case

with row type *COV* (or *CORR*) for each model parameter, plus a separate case for each of the other row types.

- **VarName_.** Takes values *P1*, *P2*, ..., corresponding to an ordered list of all estimated model parameters (except the scale or negative binomial parameters), for row types *COV* or *CORR*, with value labels corresponding to the parameter strings shown in the Parameter estimates table. The cells are blank for other row types.
- **P1, P2, ...** These variables correspond to an ordered list of all model parameters (including the scale and negative binomial parameters, as appropriate), with variable labels corresponding to the parameter strings shown in the Parameter estimates table, and take values according to the row type.

  For redundant parameters, all covariances are set to zero, correlations are set to the system-missing value; all parameter estimates are set at zero; and all standard errors, significance levels, and residual degrees of freedom are set to the system-missing value.

  For the scale parameter, covariances, correlations, significance level and degrees of freedom are set to the system-missing value. If the scale parameter is estimated via maximum likelihood, the standard error is given; otherwise it is set to the system-missing value.

  For the negative binomial parameter, covariances, correlations, significance level and degrees of freedom are set to the system-missing value. If the negative binomial parameter is estimated via maximum likelihood, the standard error is given; otherwise it is set to the system-missing value.

  If there are splits, then the list of parameters must be accumulated across all splits. In a given split, some parameters may be irrelevant; this is not the same as redundant. For irrelevant parameters, all covariances or correlations, parameter estimates, standard errors, significance levels, and degrees of freedom are set to the system-missing value.

You can use this matrix file as the initial values for further model estimation; note that this file is not immediately usable for further analyses in other procedures that read a matrix file unless those procedures accept all the row types exported here. Even then, you should take care that all parameters in this matrix file have the same meaning for the procedure reading the file.

Export model as XML. Saves the parameter estimates and the parameter covariance matrix, if selected, in XML (PMML) format. SmartScore and SPSS Statistics Server (a separate product) can use this model file to apply the model information to other data files for scoring purposes.

GENLIN Command Additional Features

The command syntax language also allows you to:

- Specify initial values for parameter estimates as a list of numbers (using the `CRITERIA` subcommand).
- Specify a fixed working correlation matrix (using the `REPEATED` subcommand).
- Fix covariates at values other than their means when computing estimated marginal means (using the `EMMEANS` subcommand).
- Specify custom polynomial contrasts for estimated marginal means (using the `EMMEANS` subcommand).
- Specify a subset of the factors for which estimated marginal means are displayed to be compared using the specified contrast type (using the `TABLES` and `COMPARE` keywords of the `EMMEANS` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter
8

Model Selection Loglinear Analysis

The Model Selection Loglinear Analysis procedure analyzes multiway crosstabulations (contingency tables). It fits hierarchical loglinear models to multidimensional crosstabulations using an iterative proportional-fitting algorithm. This procedure helps you find out which categorical variables are associated. To build models, forced entry and backward elimination methods are available. For saturated models, you can request parameter estimates and tests of partial association. A saturated model adds 0.5 to all cells.

Example. In a study of user preference for one of two laundry detergents, researchers counted people in each group, combining various categories of water softness (soft, medium, or hard), previous use of one of the brands, and washing temperature (cold or hot). They found how temperature is related to water softness and also to brand preference.

Statistics. Frequencies, residuals, parameter estimates, standard errors, confidence intervals, and tests of partial association. For custom models, plots of residuals and normal probability plots.

Data. Factor variables are categorical. All variables to be analyzed must be numeric. Categorical string variables can be recoded to numeric variables before starting the model selection analysis.

Avoid specifying many variables with many levels. Such specifications can lead to a situation where many cells have small numbers of observations, and the chi-square values may not be useful.

Related procedures. The Model Selection procedure can help identify the terms needed in the model. Then you can continue to evaluate the model using General Loglinear Analysis or Logit Loglinear Analysis. You can use Autorecode to recode