IBM SPSS Advanced Statistics 23

Note

Before using this information and the product it supports, read the information in “Notices” on page 103.

Product Information

This edition applies to version 23, release 0, modification 0 of IBM SPSS Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Chapter 1. Introduction to Advanced Statistics

The Advanced Statistics option provides procedures that offer more advanced modeling options than are available through the Statistics Base option.

- GLM Multivariate extends the general linear model provided by GLM Univariate to allow multiple dependent variables. A further extension, GLM Repeated Measures, allows repeated measurements of multiple dependent variables.
- Variance Components Analysis is a specific tool for decomposing the variability in a dependent variable into fixed and random components.
- Linear Mixed Models expands the general linear model so that the data are permitted to exhibit correlated and nonconstant variability. The mixed linear model, therefore, provides the flexibility of modeling not only the means of the data but the variances and covariances as well.
- Generalized Linear Models (GZLM) relaxes the assumption of normality for the error term and requires only that the dependent variable be linearly related to the predictors through a transformation, or link function. Generalized Estimating Equations (GEE) extends GZLM to allow repeated measurements.
- General Loglinear Analysis allows you to fit models for cross-classified count data, and Model Selection Loglinear Analysis can help you to choose between models.
- Logit Loglinear Analysis allows you to fit loglinear models for analyzing the relationship between a categorical dependent and one or more categorical predictors.
- Survival analysis is available through Life Tables for examining the distribution of time-to-event variables, possibly by levels of a factor variable; Kaplan-Meier Survival Analysis for examining the distribution of time-to-event variables, possibly by levels of a factor variable or producing separate analyses by levels of a stratification variable; and Cox Regression for modeling the time to a specified event, based upon the values of given covariates.

Chapter 2. GLM Multivariate Analysis

The GLM Multivariate procedure provides regression analysis and analysis of variance for multiple dependent variables by one or more factor variables or covariates. The factor variables divide the population into groups. Using this general linear model procedure, you can test null hypotheses about the effects of factor variables on the means of various groupings of a joint distribution of dependent variables. You can investigate interactions between factors as well as the effects of individual factors. In addition, the effects of covariates and covariate interactions with factors can be included. For regression analysis, the independent (predictor) variables are specified as covariates.

Both balanced and unbalanced models can be tested. A design is balanced if each cell in the model contains the same number of cases. In a multivariate model, the sums of squares due to the effects in the model and error sums of squares are in matrix form rather than the scalar form found in univariate analysis. These matrices are called SSCP (sums-of-squares and cross-products) matrices. If more than one dependent variable is specified, the multivariate analysis of variance using Pillai's trace, Wilks' lambda, Hotelling's trace, and Roy's largest root criterion with approximate *F* statistic are provided as well as the univariate analysis of variance for each dependent variable. In addition to testing hypotheses, GLM Multivariate produces estimates of parameters.

Commonly used *a priori* contrasts are available to perform hypothesis testing. Additionally, after an overall *F* test has shown significance, you can use post hoc tests to evaluate differences among specific means. Estimated marginal means give estimates of predicted mean values for the cells in the model, and profile plots (interaction plots) of these means allow you to visualize some of the relationships easily. The post hoc multiple comparison tests are performed for each dependent variable separately.

Residuals, predicted values, Cook's distance, and leverage values can be saved as new variables in your data file for checking assumptions. Also available are a residual SSCP matrix, which is a square matrix of sums of squares and cross-products of residuals, a residual covariance matrix, which is the residual SSCP matrix divided by the degrees of freedom of the residuals, and the residual correlation matrix, which is the standardized form of the residual covariance matrix.

WLS Weight allows you to specify a variable used to give observations different weights for a weighted least-squares (WLS) analysis, perhaps to compensate for different precision of measurement.

Example. A manufacturer of plastics measures three properties of plastic film: tear resistance, gloss, and opacity. Two rates of extrusion and two different amounts of additive are tried, and the three properties are measured under each combination of extrusion rate and additive amount. The manufacturer finds that the extrusion rate and the amount of additive individually produce significant results but that the interaction of the two factors is not significant.

Methods. Type I, Type II, Type III, and Type IV sums of squares can be used to evaluate different hypotheses. Type III is the default.

Statistics. Post hoc range tests and multiple comparisons: least significant difference, Bonferroni, Sidak, Scheffé, Ryan-Einot-Gabriel-Welsch multiple *F*, Ryan-Einot-Gabriel-Welsch multiple range, Student-Newman-Keuls, Tukey's honestly significant difference, Tukey's *b*, Duncan, Hochberg's GT2, Gabriel, Waller Duncan *t* test, Dunnett (one-sided and two-sided), Tamhane's T2, Dunnett's T3, Games-Howell, and Dunnett's *C*. Descriptive statistics: observed means, standard deviations, and counts for all of the dependent variables in all cells; the Levene test for homogeneity of variance; Box's *M* test of the homogeneity of the covariance matrices of the dependent variables; and Bartlett's test of sphericity.

Plots. Spread-versus-level, residual, and profile (interaction).

GLM Multivariate Data Considerations

Data. The dependent variables should be quantitative. Factors are categorical and can have numeric values or string values. Covariates are quantitative variables that are related to the dependent variable.

Assumptions. For dependent variables, the data are a random sample of vectors from a multivariate normal population; in the population, the variance-covariance matrices for all cells are the same. Analysis of variance is robust to departures from normality, although the data should be symmetric. To check assumptions, you can use homogeneity of variances tests (including Box's *M*) and spread-versus-level plots. You can also examine residuals and residual plots.

Related procedures. Use the Explore procedure to examine the data before doing an analysis of variance. For a single dependent variable, use GLM Univariate. If you measured the same dependent variables on several occasions for each subject, use GLM Repeated Measures.

Obtaining GLM Multivariate Tables

1. From the menus choose:
   Analyze > **General Linear Model** > **Multivariate...**
2. Select at least two dependent variables.

Optionally, you can specify Fixed Factor(s), Covariate(s), and WLS Weight.

GLM Multivariate Model

Specify Model. A full factorial model contains all factor main effects, all covariate main effects, and all factor-by-factor interactions. It does not contain covariate interactions. Select **Custom** to specify only a subset of interactions or to specify factor-by-covariate interactions. You must indicate all of the terms to be included in the model.

Factors and Covariates. The factors and covariates are listed.

Model. The model depends on the nature of your data. After selecting **Custom**, you can select the main effects and interactions that are of interest in your analysis.

Sum of squares. The method of calculating the sums of squares. For balanced or unbalanced models with no missing cells, the Type III sum-of-squares method is most commonly used.

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

Build Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term of all selected variables. This is the default.

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

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.

Sum of Squares

For the model, you can choose a type of sums of squares. Type III is the most commonly used and is the default.

Type I. This method is also known as the hierarchical decomposition of the sum-of-squares method. Each term is adjusted for only the term that precedes it in the model. Type I sums of squares are commonly used for:

- A balanced ANOVA model in which any main effects are specified before any first-order interaction effects, any first-order interaction effects are specified before any second-order interaction effects, and so on.
- A polynomial regression model in which any lower-order terms are specified before any higher-order terms.
- A purely nested model in which the first-specified effect is nested within the second-specified effect, the second-specified effect is nested within the third, and so on. (This form of nesting can be specified only by using syntax.)

Type II. This method calculates the sums of squares of an effect in the model adjusted for all other "appropriate" effects. An appropriate effect is one that corresponds to all effects that do not contain the effect being examined. The Type II sum-of-squares method is commonly used for:

- A balanced ANOVA model.
- Any model that has main factor effects only.
- Any regression model.
- A purely nested design. (This form of nesting can be specified by using syntax.)

Type III. The default. This method calculates the sums of squares of an effect in the design as the sums of squares, adjusted for any other effects that do not contain the effect, and orthogonal to any effects (if any) that contain the effect. The Type III sums of squares have one major advantage in that they are invariant with respect to the cell frequencies as long as the general form of estimability remains constant. Hence, this type of sums of squares is often considered useful for an unbalanced model with no missing cells. In a factorial design with no missing cells, this method is equivalent to the Yates' weighted-squares-of-means technique. The Type III sum-of-squares method is commonly used for:

- Any models listed in Type I and Type II.
- Any balanced or unbalanced model with no empty cells.

Type IV. This method is designed for a situation in which there are missing cells. For any effect *F* in the design, if *F* is not contained in any other effect, then Type IV = Type III = Type II. When *F* is contained in other effects, Type IV distributes the contrasts being made among the parameters in *F* to all higher-level effects equitably. The Type IV sum-of-squares method is commonly used for:

- Any models listed in Type I and Type II.
- Any balanced model or unbalanced model with empty cells.

GLM Multivariate Contrasts

Contrasts are used to test whether the levels of an effect are significantly different from one another. You can specify a contrast for each factor in the model. Contrasts represent linear combinations of the parameters.

Hypothesis testing is based on the null hypothesis **LBM = 0**, where **L** is the contrast coefficients matrix, **M** is the identity matrix (which has dimension equal to the number of dependent variables), and **B** is the parameter vector. When a contrast is specified, an **L** matrix is created such that the columns corresponding to the factor match the contrast. The remaining columns are adjusted so that the **L** matrix is estimable.

In addition to the univariate test using *F* statistics and the Bonferroni-type simultaneous confidence intervals based on Student's *t* distribution for the contrast differences across all dependent variables, the multivariate tests using Pillai's trace, Wilks' lambda, Hotelling's trace, and Roy's largest root criteria are provided.

Available contrasts are deviation, simple, difference, Helmert, repeated, and polynomial. For deviation contrasts and simple contrasts, you can choose whether the reference category is the last or first category.

Contrast Types

Deviation. Compares the mean of each level (except a reference category) to the mean of all of the levels (grand mean). The levels of the factor can be in any order.

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. You can choose the first or last category as the reference.

Difference. Compares the mean of each level (except the first) to the mean of previous levels. (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.

GLM Multivariate Profile Plots

Profile plots (interaction plots) are useful for comparing marginal means in your model. A profile plot is a line plot in which each point indicates the estimated marginal mean of a dependent variable (adjusted for any covariates) at one level of a factor. The levels of a second factor can be used to make separate lines. Each level in a third factor can be used to create a separate plot. All factors are available for plots. Profile plots are created for each dependent variable.

A profile plot of one factor shows whether the estimated marginal means are increasing or decreasing across levels. For two or more factors, parallel lines indicate that there is no interaction between factors, which means that you can investigate the levels of only one factor. Nonparallel lines indicate an interaction.

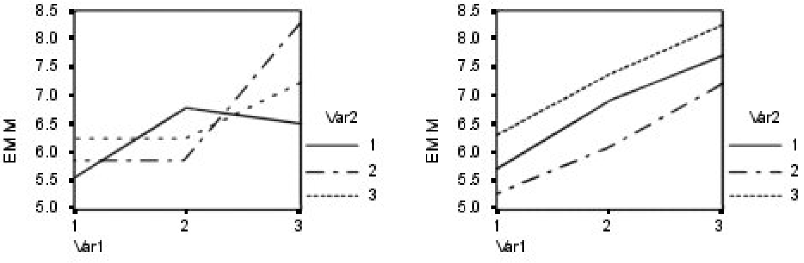

Figure 1. Nonparallel plot (left) and parallel plot (right)

After a plot is specified by selecting factors for the horizontal axis and, optionally, factors for separate lines and separate plots, the plot must be added to the Plots list.

GLM Multivariate Post Hoc Comparisons

Post hoc multiple comparison tests. Once you have determined that differences exist among the means, post hoc range tests and pairwise multiple comparisons can determine which means differ. Comparisons are made on unadjusted values. The post hoc tests are performed for each dependent variable separately.

The Bonferroni and Tukey's honestly significant difference tests are commonly used multiple comparison tests. The **Bonferroni test**, based on Student's *t* statistic, adjusts the observed significance level for the fact that multiple comparisons are made. **Sidak's t test** also adjusts the significance level and provides tighter bounds than the Bonferroni test. **Tukey's honestly significant difference test** uses the Studentized range statistic to make all pairwise comparisons between groups and sets the experimentwise error rate to the error rate for the collection for all pairwise comparisons. When testing a large number of pairs of means, Tukey's honestly significant difference test is more powerful than the Bonferroni test. For a small number of pairs, Bonferroni is more powerful.

Hochberg's GT2 is similar to Tukey's honestly significant difference test, but the Studentized maximum modulus is used. Usually, Tukey's test is more powerful. **Gabriel's pairwise comparisons test** also uses the Studentized maximum modulus and is generally more powerful than Hochberg's GT2 when the cell sizes are unequal. Gabriel's test may become liberal when the cell sizes vary greatly.

Dunnett's pairwise multiple comparison t test compares a set of treatments against a single control mean. The last category is the default control category. Alternatively, you can choose the first category. You can also choose a two-sided or one-sided test. To test that the mean at any level (except the control category) of the factor is not equal to that of the control category, use a two-sided test. To test whether the mean at any level of the factor is smaller than that of the control category, select **< Control**. Likewise, to test whether the mean at any level of the factor is larger than that of the control category, select **> Control**.

Ryan, Einot, Gabriel, and Welsch (R-E-G-W) developed two multiple step-down range tests. Multiple step-down procedures first test whether all means are equal. If all means are not equal, subsets of means are tested for equality. **R-E-G-W F** is based on an *F* test and **R-E-G-W Q** is based on the Studentized range. These tests are more powerful than Duncan's multiple range test and Student-Newman-Keuls (which are also multiple step-down procedures), but they are not recommended for unequal cell sizes.

When the variances are unequal, use **Tamhane's T2** (conservative pairwise comparisons test based on a *t* test), **Dunnett's T3** (pairwise comparison test based on the Studentized maximum modulus), **Games-Howell pairwise comparison test** (sometimes liberal), or **Dunnett's C** (pairwise comparison test based on the Studentized range).

Duncan's multiple range test, Student-Newman-Keuls (**S-N-K**), and **Tukey's b** are range tests that rank group means and compute a range value. These tests are not used as frequently as the tests previously discussed.

The **Waller-Duncan t test** uses a Bayesian approach. This range test uses the harmonic mean of the sample size when the sample sizes are unequal.

The significance level of the **Scheffé** test is designed to allow all possible linear combinations of group means to be tested, not just pairwise comparisons available in this feature. The result is that the Scheffé test is often more conservative than other tests, which means that a larger difference between means is required for significance.

The least significant difference (**LSD**) pairwise multiple comparison test is equivalent to multiple individual *t* tests between all pairs of groups. The disadvantage of this test is that no attempt is made to adjust the observed significance level for multiple comparisons.

Tests displayed. Pairwise comparisons are provided for LSD, Sidak, Bonferroni, Games-Howell, Tamhane's T2 and T3, Dunnett's *C*, and Dunnett's T3. Homogeneous subsets for range tests are provided for S-N-K, Tukey's *b*, Duncan, R-E-G-W *F*, R-E-G-W *Q*, and Waller. Tukey's honestly significant difference test, Hochberg's GT2, Gabriel's test, and Scheffé's test are both multiple comparison tests and range tests.

GLM Save

You can save values predicted by the model, residuals, and related measures as new variables in the Data Editor. Many of these variables can be used for examining assumptions about the data. To save the values for use in another IBM® SPSS® Statistics session, you must save the current data file.

Predicted Values. The values that the model predicts for each case.

- *Unstandardized*. The value the model predicts for the dependent variable.
- *Weighted*. Weighted unstandardized predicted values. Available only if a WLS variable was previously selected.
- *Standard error*. An estimate of the standard deviation of the average value of the dependent variable for cases that have the same values of the independent variables.

Diagnostics. Measures to identify cases with unusual combinations of values for the independent variables and cases that may have a large impact on the model.

- *Cook's distance*. A measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients. A large Cook's D indicates that excluding a case from computation of the regression statistics changes the coefficients substantially.
- *Leverage values*. Uncentered leverage values. The relative influence of each observation on the model's fit.

Residuals. An unstandardized residual is the actual value of the dependent variable minus the value predicted by the model. Standardized, Studentized, and deleted residuals are also available. If a WLS variable was chosen, weighted unstandardized residuals are available.

- *Unstandardized*. The difference between an observed value and the value predicted by the model.
- *Weighted*. Weighted unstandardized residuals. Available only if a WLS variable was previously selected.
- *Standardized*. The residual divided by an estimate of its standard deviation. Standardized residuals, which are also known as Pearson residuals, have a mean of 0 and a standard deviation of 1.
- *Studentized*. The residual divided by an estimate of its standard deviation that varies from case to case, depending on the distance of each case's values on the independent variables from the means of the independent variables.
- *Deleted*. The residual for a case when that case is excluded from the calculation of the regression coefficients. It is the difference between the value of the dependent variable and the adjusted predicted value.

Coefficient Statistics. Writes a variance-covariance matrix of the parameter estimates in the model to a new dataset in the current session or an external IBM SPSS Statistics data file. Also, for each dependent variable, there will be a row of parameter estimates, a row of significance values for the *t* statistics corresponding to the parameter estimates, and a row of residual degrees of freedom. For a multivariate model, there are similar rows for each dependent variable. You can use this matrix file in other procedures that read matrix files.

GLM Multivariate Options

Optional statistics are available from this dialog box. Statistics are calculated using a fixed-effects model.

Estimated Marginal Means. Select the factors and interactions for which you want estimates of the population marginal means in the cells. These means are adjusted for the covariates, if any. Interactions are available only if you have specified a custom model.

- **Compare main effects.** Provides uncorrected pairwise comparisons among estimated marginal means for any main effect in the model, for both between- and within-subjects factors. This item is available only if main effects are selected under the Display Means For list.
- **Confidence interval adjustment.** Select least significant difference (LSD), Bonferroni, or Sidak adjustment to the confidence intervals and significance. This item is available only if **Compare main effects** is selected.

Display. Select **Descriptive statistics** to produce observed means, standard deviations, and counts for all of the dependent variables in all cells. **Estimates of effect size** gives a partial eta-squared value for each effect and each parameter estimate. The eta-squared statistic describes the proportion of total variability attributable to a factor. Select **Observed power** to obtain the power of the test when the alternative hypothesis is set based on the observed value. Select **Parameter estimates** to produce the parameter estimates, standard errors, *t* tests, confidence intervals, and the observed power for each test. You can display the hypothesis and error **SSCP matrices** and the **Residual SSCP matrix** plus Bartlett's test of sphericity of the residual covariance matrix.

Homogeneity tests produces the Levene test of the homogeneity of variance for each dependent variable across all level combinations of the between-subjects factors, for between-subjects factors only. Also, homogeneity tests include Box's *M* test of the homogeneity of the covariance matrices of the dependent variables across all level combinations of the between-subjects factors. The spread-versus-level and residual plots options are useful for checking assumptions about the data. This item is disabled if there are no factors. Select **Residual plots** to produce an observed-by-predicted-by-standardized residuals plot for each dependent variable. These plots are useful for investigating the assumption of equal variance. Select **Lack of fit test** to check if the relationship between the dependent variable and the independent variables can be adequately described by the model. **General estimable function** allows you to construct custom hypothesis tests based on the general estimable function. Rows in any contrast coefficient matrix are linear combinations of the general estimable function.

Significance level. You might want to adjust the significance level used in post hoc tests and the confidence level used for constructing confidence intervals. The specified value is also used to calculate the observed power for the test. When you specify a significance level, the associated level of the confidence intervals is displayed in the dialog box.

GLM Command Additional Features

These features may apply to univariate, multivariate, or repeated measures analysis. The command syntax language also allows you to:

- Specify nested effects in the design (using the `DESIGN` subcommand).
- Specify tests of effects versus a linear combination of effects or a value (using the `TEST` subcommand).
- Specify multiple contrasts (using the `CONTRAST` subcommand).
- Include user-missing values (using the `MISSING` subcommand).
- Specify EPS criteria (using the `CRITERIA` subcommand).
- Construct a custom **L** matrix, **M** matrix, or **K** matrix (using the `LMATRIX`, `MMATRIX`, or `KMATRIX` subcommands).
- For deviation or simple contrasts, specify an intermediate reference category (using the `CONTRAST` subcommand).
- Specify metrics for polynomial contrasts (using the `CONTRAST` subcommand).
- Specify error terms for post hoc comparisons (using the `POSTHOC` subcommand).
- Compute estimated marginal means for any factor or factor interaction among the factors in the factor list (using the `EMMEANS` subcommand).

- Specify names for temporary variables (using the `SAVE` subcommand).
- Construct a correlation matrix data file (using the `OUTFILE` subcommand).
- Construct a matrix data file that contains statistics from the between-subjects ANOVA table (using the `OUTFILE` subcommand).
- Save the design matrix to a new data file (using the `OUTFILE` subcommand).

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

Chapter 3. GLM Repeated Measures

The GLM Repeated Measures procedure provides analysis of variance when the same measurement is made several times on each subject or case. If between-subjects factors are specified, they divide the population into groups. Using this general linear model procedure, you can test null hypotheses about the effects of both the between-subjects factors and the within-subjects factors. You can investigate interactions between factors as well as the effects of individual factors. In addition, the effects of constant covariates and covariate interactions with the between-subjects factors can be included.

In a doubly multivariate repeated measures design, the dependent variables represent measurements of more than one variable for the different levels of the within-subjects factors. For example, you could have measured both pulse and respiration at three different times on each subject.

The GLM Repeated Measures procedure provides both univariate and multivariate analyses for the repeated measures data. Both balanced and unbalanced models can be tested. A design is balanced if each cell in the model contains the same number of cases. In a multivariate model, the sums of squares due to the effects in the model and error sums of squares are in matrix form rather than the scalar form found in univariate analysis. These matrices are called SSCP (sums-of-squares and cross-products) matrices. In addition to testing hypotheses, GLM Repeated Measures produces estimates of parameters.

Commonly used *a priori* contrasts are available to perform hypothesis testing on between-subjects factors. Additionally, after an overall *F* test has shown significance, you can use post hoc tests to evaluate differences among specific means. Estimated marginal means give estimates of predicted mean values for the cells in the model, and profile plots (interaction plots) of these means allow you to visualize some of the relationships easily.

Residuals, predicted values, Cook's distance, and leverage values can be saved as new variables in your data file for checking assumptions. Also available are a residual SSCP matrix, which is a square matrix of sums of squares and cross-products of residuals, a residual covariance matrix, which is the residual SSCP matrix divided by the degrees of freedom of the residuals, and the residual correlation matrix, which is the standardized form of the residual covariance matrix.

WLS Weight allows you to specify a variable used to give observations different weights for a weighted least-squares (WLS) analysis, perhaps to compensate for different precision of measurement.

Example. Twelve students are assigned to a high- or low-anxiety group based on their scores on an anxiety-rating test. The anxiety rating is called a between-subjects factor because it divides the subjects into groups. The students are each given four trials on a learning task, and the number of errors for each trial is recorded. The errors for each trial are recorded in separate variables, and a within-subjects factor (trial) is defined with four levels for the four trials. The trial effect is found to be significant, while the trial-by-anxiety interaction is not significant.

Methods. Type I, Type II, Type III, and Type IV sums of squares can be used to evaluate different hypotheses. Type III is the default.

Statistics. Post hoc range tests and multiple comparisons (for between-subjects factors): least significant difference, Bonferroni, Sidak, Scheffé, Ryan-Einot-Gabriel-Welsch multiple *F*, Ryan-Einot-Gabriel-Welsch multiple range, Student-Newman-Keuls, Tukey's honestly significant difference, Tukey's *b*, Duncan, Hochberg's GT2, Gabriel, Waller Duncan *t* test, Dunnett (one-sided and two-sided), Tamhane's T2, Dunnett's T3, Games-Howell, and Dunnett's *C*. Descriptive statistics: observed means, standard deviations, and counts for all of the dependent variables in all cells; the Levene test for homogeneity of variance; Box's *M*; and Mauchly's test of sphericity.

Plots. Spread-versus-level, residual, and profile (interaction).

GLM Repeated Measures Data Considerations

Data. The dependent variables should be quantitative. Between-subjects factors divide the sample into discrete subgroups, such as male and female. These factors are categorical and can have numeric values or string values. Within-subjects factors are defined in the Repeated Measures Define Factor(s) dialog box. Covariates are quantitative variables that are related to the dependent variable. For a repeated measures analysis, these should remain constant at each level of a within-subjects variable.

The data file should contain a set of variables for each group of measurements on the subjects. The set has one variable for each repetition of the measurement within the group. A within-subjects factor is defined for the group with the number of levels equal to the number of repetitions. For example, measurements of weight could be taken on different days. If measurements of the same property were taken on five days, the within-subjects factor could be specified as *day* with five levels.

For multiple within-subjects factors, the number of measurements for each subject is equal to the product of the number of levels of each factor. For example, if measurements were taken at three different times each day for four days, the total number of measurements is 12 for each subject. The within-subjects factors could be specified as *day(4)* and *time(3)*.

Assumptions. A repeated measures analysis can be approached in two ways, univariate and multivariate.

The univariate approach (also known as the split-plot or mixed-model approach) considers the dependent variables as responses to the levels of within-subjects factors. The measurements on a subject should be a sample from a multivariate normal distribution, and the variance-covariance matrices are the same across the cells formed by the between-subjects effects. Certain assumptions are made on the variance-covariance matrix of the dependent variables. The validity of the *F* statistic used in the univariate approach can be assured if the variance-covariance matrix is circular in form (Huynh and Mandeville, 1979).

To test this assumption, Mauchly's test of sphericity can be used, which performs a test of sphericity on the variance-covariance matrix of an orthonormalized transformed dependent variable. Mauchly's test is automatically displayed for a repeated measures analysis. For small sample sizes, this test is not very powerful. For large sample sizes, the test may be significant even when the impact of the departure on the results is small. If the significance of the test is large, the hypothesis of sphericity can be assumed. However, if the significance is small and the sphericity assumption appears to be violated, an adjustment to the numerator and denominator degrees of freedom can be made in order to validate the univariate *F* statistic. Three estimates of this adjustment, which is called **epsilon**, are available in the GLM Repeated Measures procedure. Both the numerator and denominator degrees of freedom must be multiplied by epsilon, and the significance of the *F* ratio must be evaluated with the new degrees of freedom.

The multivariate approach considers the measurements on a subject to be a sample from a multivariate normal distribution, and the variance-covariance matrices are the same across the cells formed by the between-subjects effects. To test whether the variance-covariance matrices across the cells are the same, Box's *M* test can be used.

Related procedures. Use the Explore procedure to examine the data before doing an analysis of variance. If there are *not* repeated measurements on each subject, use GLM Univariate or GLM Multivariate. If there are only two measurements for each subject (for example, pre-test and post-test) and there are no between-subjects factors, you can use the Paired-Samples T Test procedure.

Obtaining GLM Repeated Measures

1. From the menus choose:

   Analyze > **General Linear Model** > **Repeated Measures...**

2. Type a within-subject factor name and its number of levels.
3. Click **Add**.
4. Repeat these steps for each within-subjects factor.

   To define measure factors for a doubly multivariate repeated measures design:
5. Type the measure name.
6. Click **Add**.

   After defining all of your factors and measures:
7. Click **Define**.
8. Select a dependent variable that corresponds to each combination of within-subjects factors (and optionally, measures) on the list.

To change positions of the variables, use the up and down arrows.

To make changes to the within-subjects factors, you can reopen the Repeated Measures Define Factor(s) dialog box without closing the main dialog box. Optionally, you can specify between-subjects factor(s) and covariates.

GLM Repeated Measures Define Factors

GLM Repeated Measures analyzes groups of related dependent variables that represent different measurements of the same attribute. This dialog box lets you define one or more within-subjects factors for use in GLM Repeated Measures. Note that the order in which you specify within-subjects factors is important. Each factor constitutes a level within the previous factor.

To use Repeated Measures, you must set up your data correctly. You must define within-subjects factors in this dialog box. Notice that these factors are not existing variables in your data but rather factors that you define here.

Example. In a weight-loss study, suppose the weights of several people are measured each week for five weeks. In the data file, each person is a subject or case. The weights for the weeks are recorded in the variables *weight1*, *weight2*, and so on. The gender of each person is recorded in another variable. The weights, measured for each subject repeatedly, can be grouped by defining a within-subjects factor. The factor could be called *week*, defined to have five levels. In the main dialog box, the variables *weight1*, ..., *weight5* are used to assign the five levels of *week*. The variable in the data file that groups males and females (*gender*) can be specified as a between-subjects factor to study the differences between males and females.

Measures. If subjects were tested on more than one measure at each time, define the measures. For example, the pulse and respiration rate could be measured on each subject every day for a week. These measures do not exist as variables in the data file but are defined here. A model with more than one measure is sometimes called a doubly multivariate repeated measures model.

GLM Repeated Measures Model

Specify Model. A full factorial model contains all factor main effects, all covariate main effects, and all factor-by-factor interactions. It does not contain covariate interactions. Select **Custom** to specify only a subset of interactions or to specify factor-by-covariate interactions. You must indicate all of the terms to be included in the model.

Between-Subjects. The between-subjects factors and covariates are listed.

Model. The model depends on the nature of your data. After selecting **Custom**, you can select the within-subjects effects and interactions and the between-subjects effects and interactions that are of interest in your analysis.

Sum of squares. The method of calculating the sums of squares for the between-subjects model. For balanced or unbalanced between-subjects models with no missing cells, the Type III sum-of-squares method is the most commonly used.

Build Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term of all selected variables. This is the default.

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

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.

Sum of Squares

For the model, you can choose a type of sums of squares. Type III is the most commonly used and is the default.

Type I. This method is also known as the hierarchical decomposition of the sum-of-squares method. Each term is adjusted for only the term that precedes it in the model. Type I sums of squares are commonly used for:

- A balanced ANOVA model in which any main effects are specified before any first-order interaction effects, any first-order interaction effects are specified before any second-order interaction effects, and so on.
- A polynomial regression model in which any lower-order terms are specified before any higher-order terms.
- A purely nested model in which the first-specified effect is nested within the second-specified effect, the second-specified effect is nested within the third, and so on. (This form of nesting can be specified only by using syntax.)

Type II. This method calculates the sums of squares of an effect in the model adjusted for all other "appropriate" effects. An appropriate effect is one that corresponds to all effects that do not contain the effect being examined. The Type II sum-of-squares method is commonly used for:

- A balanced ANOVA model.
- Any model that has main factor effects only.
- Any regression model.
- A purely nested design. (This form of nesting can be specified by using syntax.)

Type III. The default. This method calculates the sums of squares of an effect in the design as the sums of squares, adjusted for any other effects that do not contain the effect, and orthogonal to any effects (if any) that contain the effect. The Type III sums of squares have one major advantage in that they are invariant with respect to the cell frequencies as long as the general form of estimability remains constant. Hence, this type of sums of squares is often considered useful for an unbalanced model with no missing cells. In a factorial design with no missing cells, this method is equivalent to the Yates' weighted-squares-of-means technique. The Type III sum-of-squares method is commonly used for:

- Any models listed in Type I and Type II.
- Any balanced or unbalanced model with no empty cells.

Type IV. This method is designed for a situation in which there are missing cells. For any effect *F* in the design, if *F* is not contained in any other effect, then Type IV = Type III = Type II. When *F* is contained in other effects, Type IV distributes the contrasts being made among the parameters in *F* to all higher-level effects equitably. The Type IV sum-of-squares method is commonly used for:

- Any models listed in Type I and Type II.
- Any balanced model or unbalanced model with empty cells.

GLM Repeated Measures Contrasts

Contrasts are used to test for differences among the levels of a between-subjects factor. You can specify a contrast for each between-subjects factor in the model. Contrasts represent linear combinations of the parameters.

Hypothesis testing is based on the null hypothesis **LBM**=0, where **L** is the contrast coefficients matrix, **B** is the parameter vector, and **M** is the average matrix that corresponds to the average transformation for the dependent variable. You can display this transformation matrix by selecting **Transformation matrix** in the Repeated Measures Options dialog box. For example, if there are four dependent variables, a within-subjects factor of four levels, and polynomial contrasts (the default) are used for within-subjects factors, the **M** matrix will be (0.5 0.5 0.5 0.5)'. When a contrast is specified, an **L** matrix is created such that the columns corresponding to the between-subjects factor match the contrast. The remaining columns are adjusted so that the **L** matrix is estimable.

Available contrasts are deviation, simple, difference, Helmert, repeated, and polynomial. For deviation contrasts and simple contrasts, you can choose whether the reference category is the last or first category.

A contrast other than **None** must be selected for within-subjects factors.

Contrast Types

Deviation. Compares the mean of each level (except a reference category) to the mean of all of the levels (grand mean). The levels of the factor can be in any order.

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. You can choose the first or last category as the reference.

Difference. Compares the mean of each level (except the first) to the mean of previous levels. (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.

GLM Repeated Measures Profile Plots

Profile plots (interaction plots) are useful for comparing marginal means in your model. A profile plot is a line plot in which each point indicates the estimated marginal mean of a dependent variable (adjusted for any covariates) at one level of a factor. The levels of a second factor can be used to make separate lines. Each level in a third factor can be used to create a separate plot. All factors are available for plots. Profile plots are created for each dependent variable. Both between-subjects factors and within-subjects factors can be used in profile plots.

A profile plot of one factor shows whether the estimated marginal means are increasing or decreasing across levels. For two or more factors, parallel lines indicate that there is no interaction between factors, which means that you can investigate the levels of only one factor. Nonparallel lines indicate an interaction.

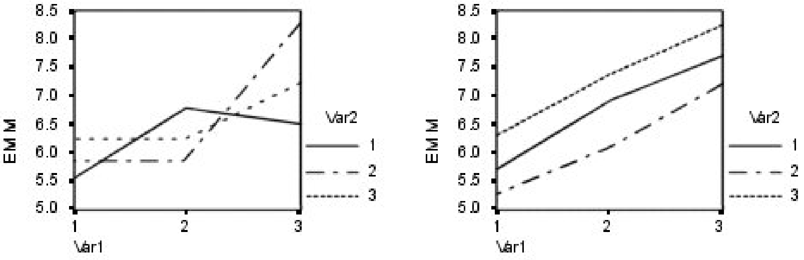

Figure 2. Nonparallel plot (left) and parallel plot (right)

After a plot is specified by selecting factors for the horizontal axis and, optionally, factors for separate lines and separate plots, the plot must be added to the Plots list.

GLM Repeated Measures Post Hoc Comparisons

Post hoc multiple comparison tests. Once you have determined that differences exist among the means, post hoc range tests and pairwise multiple comparisons can determine which means differ. Comparisons are made on unadjusted values. These tests are not available if there are no between-subjects factors, and the post hoc multiple comparison tests are performed for the average across the levels of the within-subjects factors.

The Bonferroni and Tukey's honestly significant difference tests are commonly used multiple comparison tests. The **Bonferroni test**, based on Student's *t* statistic, adjusts the observed significance level for the fact that multiple comparisons are made. **Sidak's t test** also adjusts the significance level and provides tighter bounds than the Bonferroni test. **Tukey's honestly significant difference test** uses the Studentized range statistic to make all pairwise comparisons between groups and sets the experimentwise error rate to the error rate for the collection for all pairwise comparisons. When testing a large number of pairs of means, Tukey's honestly significant difference test is more powerful than the Bonferroni test. For a small number of pairs, Bonferroni is more powerful.

Hochberg's GT2 is similar to Tukey's honestly significant difference test, but the Studentized maximum modulus is used. Usually, Tukey's test is more powerful. **Gabriel's pairwise comparisons test** also uses the Studentized maximum modulus and is generally more powerful than Hochberg's GT2 when the cell sizes are unequal. Gabriel's test may become liberal when the cell sizes vary greatly.

Dunnett's pairwise multiple comparison t test compares a set of treatments against a single control mean. The last category is the default control category. Alternatively, you can choose the first category. You can also choose a two-sided or one-sided test. To test that the mean at any level (except the control category) of the factor is not equal to that of the control category, use a two-sided test. To test whether the mean at any level of the factor is smaller than that of the control category, select **< Control**. Likewise, to test whether the mean at any level of the factor is larger than that of the control category, select **> Control**.

Ryan, Einot, Gabriel, and Welsch (R-E-G-W) developed two multiple step-down range tests. Multiple step-down procedures first test whether all means are equal. If all means are not equal, subsets of means are tested for equality. **R-E-G-W F** is based on an *F* test and **R-E-G-W Q** is based on the Studentized range. These tests are more powerful than Duncan's multiple range test and Student-Newman-Keuls (which are also multiple step-down procedures), but they are not recommended for unequal cell sizes.

When the variances are unequal, use **Tamhane's T2** (conservative pairwise comparisons test based on a *t* test), **Dunnett's T3** (pairwise comparison test based on the Studentized maximum modulus), **Games-Howell pairwise comparison test** (sometimes liberal), or **Dunnett's C** (pairwise comparison test based on the Studentized range).

Duncan's multiple range test, Student-Newman-Keuls (**S-N-K**), and **Tukey's b** are range tests that rank group means and compute a range value. These tests are not used as frequently as the tests previously discussed.

The **Waller-Duncan t test** uses a Bayesian approach. This range test uses the harmonic mean of the sample size when the sample sizes are unequal.

The significance level of the **Scheffé** test is designed to allow all possible linear combinations of group means to be tested, not just pairwise comparisons available in this feature. The result is that the Scheffé test is often more conservative than other tests, which means that a larger difference between means is required for significance.

The least significant difference (**LSD**) pairwise multiple comparison test is equivalent to multiple individual *t* tests between all pairs of groups. The disadvantage of this test is that no attempt is made to adjust the observed significance level for multiple comparisons.

Tests displayed. Pairwise comparisons are provided for LSD, Sidak, Bonferroni, Games-Howell, Tamhane's T2 and T3, Dunnett's *C*, and Dunnett's T3. Homogeneous subsets for range tests are provided for S-N-K, Tukey's *b*, Duncan, R-E-G-W *F*, R-E-G-W *Q*, and Waller. Tukey's honestly significant difference test, Hochberg's GT2, Gabriel's test, and Scheffé's test are both multiple comparison tests and range tests.

GLM Repeated Measures Save

You can save values predicted by the model, residuals, and related measures as new variables in the Data Editor. Many of these variables can be used for examining assumptions about the data. To save the values for use in another IBM SPSS Statistics session, you must save the current data file.

Predicted Values. The values that the model predicts for each case.

- *Unstandardized*. The value the model predicts for the dependent variable.
- *Standard error*. An estimate of the standard deviation of the average value of the dependent variable for cases that have the same values of the independent variables.

Diagnostics. Measures to identify cases with unusual combinations of values for the independent variables and cases that may have a large impact on the model. Available are Cook's distance and uncentered leverage values.

- *Cook's distance*. A measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients. A large Cook's D indicates that excluding a case from computation of the regression statistics changes the coefficients substantially.
- *Leverage values*. Uncentered leverage values. The relative influence of each observation on the model's fit.

Residuals. An unstandardized residual is the actual value of the dependent variable minus the value predicted by the model. Standardized, Studentized, and deleted residuals are also available.

- *Unstandardized*. The difference between an observed value and the value predicted by the model.
- *Standardized*. The residual divided by an estimate of its standard deviation. Standardized residuals, which are also known as Pearson residuals, have a mean of 0 and a standard deviation of 1.
- *Studentized*. The residual divided by an estimate of its standard deviation that varies from case to case, depending on the distance of each case's values on the independent variables from the means of the independent variables.

- *Deleted*. The residual for a case when that case is excluded from the calculation of the regression coefficients. It is the difference between the value of the dependent variable and the adjusted predicted value.

Coefficient Statistics. Saves a variance-covariance matrix of the parameter estimates to a dataset or a data file. Also, for each dependent variable, there will be a row of parameter estimates, a row of significance values for the *t* statistics corresponding to the parameter estimates, and a row of residual degrees of freedom. For a multivariate model, there are similar rows for each dependent variable. You can use this matrix data in other procedures that read matrix files. Datasets are available for subsequent use in the same session but are not saved as files unless explicitly saved prior to the end of the session. Dataset names must conform to variable naming rules.

GLM Repeated Measures Options

Optional statistics are available from this dialog box. Statistics are calculated using a fixed-effects model.

Estimated Marginal Means. Select the factors and interactions for which you want estimates of the population marginal means in the cells. These means are adjusted for the covariates, if any. Both within-subjects and between-subjects factors can be selected.

- **Compare main effects.** Provides uncorrected pairwise comparisons among estimated marginal means for any main effect in the model, for both between- and within-subjects factors. This item is available only if main effects are selected under the Display Means For list.
- **Confidence interval adjustment.** Select least significant difference (LSD), Bonferroni, or Sidak adjustment to the confidence intervals and significance. This item is available only if **Compare main effects** is selected.

Display. Select **Descriptive statistics** to produce observed means, standard deviations, and counts for all of the dependent variables in all cells. **Estimates of effect size** gives a partial eta-squared value for each effect and each parameter estimate. The eta-squared statistic describes the proportion of total variability attributable to a factor. Select **Observed power** to obtain the power of the test when the alternative hypothesis is set based on the observed value. Select **Parameter estimates** to produce the parameter estimates, standard errors, *t* tests, confidence intervals, and the observed power for each test. You can display the hypothesis and error **SSCP matrices** and the **Residual SSCP matrix** plus Bartlett's test of sphericity of the residual covariance matrix.

Homogeneity tests produces the Levene test of the homogeneity of variance for each dependent variable across all level combinations of the between-subjects factors, for between-subjects factors only. Also, homogeneity tests include Box's *M* test of the homogeneity of the covariance matrices of the dependent variables across all level combinations of the between-subjects factors. The spread-versus-level and residual plots options are useful for checking assumptions about the data. This item is disabled if there are no factors. Select **Residual plots** to produce an observed-by-predicted-by-standardized residuals plot for each dependent variable. These plots are useful for investigating the assumption of equal variance. Select **Lack of fit test** to check if the relationship between the dependent variable and the independent variables can be adequately described by the model. **General estimable function** allows you to construct custom hypothesis tests based on the general estimable function. Rows in any contrast coefficient matrix are linear combinations of the general estimable function.

Significance level. You might want to adjust the significance level used in post hoc tests and the confidence level used for constructing confidence intervals. The specified value is also used to calculate the observed power for the test. When you specify a significance level, the associated level of the confidence intervals is displayed in the dialog box.

GLM Command Additional Features

These features may apply to univariate, multivariate, or repeated measures analysis. The command syntax language also allows you to:

- Specify nested effects in the design (using the DESIGN subcommand).
- Specify tests of effects versus a linear combination of effects or a value (using the TEST subcommand).
- Specify multiple contrasts (using the CONTRAST subcommand).
- Include user-missing values (using the MISSING subcommand).
- Specify EPS criteria (using the CRITERIA subcommand).
- Construct a custom **L** matrix, **M** matrix, or **K** matrix (using the LMATRIX, MMATRIX, and KMATRIX subcommands).
- For deviation or simple contrasts, specify an intermediate reference category (using the CONTRAST subcommand).
- Specify metrics for polynomial contrasts (using the CONTRAST subcommand).
- Specify error terms for post hoc comparisons (using the POSTHOC subcommand).
- Compute estimated marginal means for any factor or factor interaction among the factors in the factor list (using the EMMEANS subcommand).
- Specify names for temporary variables (using the SAVE subcommand).
- Construct a correlation matrix data file (using the OUTFILE subcommand).
- Construct a matrix data file that contains statistics from the between-subjects ANOVA table (using the OUTFILE subcommand).
- Save the design matrix to a new data file (using the OUTFILE subcommand).

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

Chapter 4. Variance Components Analysis

The Variance Components procedure, for mixed-effects models, estimates the contribution of each random effect to the variance of the dependent variable. This procedure is particularly interesting for analysis of mixed models such as split plot, univariate repeated measures, and random block designs. By calculating variance components, you can determine where to focus attention in order to reduce the variance.

Four different methods are available for estimating the variance components: minimum norm quadratic unbiased estimator (MINQUE), analysis of variance (ANOVA), maximum likelihood (ML), and restricted maximum likelihood (REML). Various specifications are available for the different methods.

Default output for all methods includes variance component estimates. If the ML method or the REML method is used, an asymptotic covariance matrix table is also displayed. Other available output includes an ANOVA table and expected mean squares for the ANOVA method and an iteration history for the ML and REML methods. The Variance Components procedure is fully compatible with the GLM Univariate procedure.

WLS Weight allows you to specify a variable used to give observations different weights for a weighted analysis, perhaps to compensate for variations in precision of measurement.

Example. At an agriculture school, weight gains for pigs in six different litters are measured after one month. The litter variable is a random factor with six levels. (The six litters studied are a random sample from a large population of pig litters.) The investigator finds out that the variance in weight gain is attributable to the difference in litters much more than to the difference in pigs within a litter.

Variance Components Data Considerations

Data. The dependent variable is quantitative. Factors are categorical. They can have numeric values or string values of up to eight bytes. At least one of the factors must be random. That is, the levels of the factor must be a random sample of possible levels. Covariates are quantitative variables that are related to the dependent variable.

Assumptions. All methods assume that model parameters of a random effect have zero means and finite constant variances and are mutually uncorrelated. Model parameters from different random effects are also uncorrelated.

The residual term also has a zero mean and finite constant variance. It is uncorrelated with model parameters of any random effect. Residual terms from different observations are assumed to be uncorrelated.

Based on these assumptions, observations from the same level of a random factor are correlated. This fact distinguishes a variance component model from a general linear model.

ANOVA and MINQUE do not require normality assumptions. They are both robust to moderate departures from the normality assumption.

ML and REML require the model parameter and the residual term to be normally distributed.

Related procedures. Use the Explore procedure to examine the data before doing variance components analysis. For hypothesis testing, use GLM Univariate, GLM Multivariate, and GLM Repeated Measures.

Obtaining Variance Components Tables

1. From the menus choose:

   Analyze > **General Linear Model** > **Variance Components...**
2. Select a dependent variable.
3. Select variables for Fixed Factor(s), Random Factor(s), and Covariate(s), as appropriate for your data. For specifying a weight variable, use WLS Weight.

Variance Components Model

Specify Model. A full factorial model contains all factor main effects, all covariate main effects, and all factor-by-factor interactions. It does not contain covariate interactions. Select **Custom** to specify only a subset of interactions or to specify factor-by-covariate interactions. You must indicate all of the terms to be included in the model.

Factors & Covariates. The factors and covariates are listed.

Model. The model depends on the nature of your data. After selecting **Custom**, you can select the main effects and interactions that are of interest in your analysis. The model must contain a random factor.

Include intercept in model. Usually the intercept is included in the model. If you can assume that the data pass through the origin, you can exclude the intercept.

Build Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term of all selected variables. This is the default.

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

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.

Variance Components Options

Method. You can choose one of four methods to estimate the variance components.

- **MINQUE** (minimum norm quadratic unbiased estimator) produces estimates that are invariant with respect to the fixed effects. If the data are normally distributed and the estimates are correct, this method produces the least variance among all unbiased estimators. You can choose a method for random-effect prior weights.
- **ANOVA** (analysis of variance) computes unbiased estimates using either the Type I or Type III sums of squares for each effect. The ANOVA method sometimes produces negative variance estimates, which can indicate an incorrect model, an inappropriate estimation method, or a need for more data.
- **Maximum likelihood** (ML) produces estimates that would be most consistent with the data actually observed, using iterations. These estimates can be biased. This method is asymptotically normal. ML and REML estimates are invariant under translation. This method does not take into account the degrees of freedom used to estimate the fixed effects.
- **Restricted maximum likelihood** (REML) estimates reduce the ANOVA estimates for many (if not all) cases of balanced data. Because this method is adjusted for the fixed effects, it should have smaller standard errors than the ML method. This method takes into account the degrees of freedom used to estimate the fixed effects.

Random Effect Priors. Uniform implies that all random effects and the residual term have an equal impact on the observations. The **Zero** scheme is equivalent to assuming zero random-effect variances. Available only for the MINQUE method.

Sum of Squares. Type I sums of squares are used for the hierarchical model, which is often used in variance component literature. If you choose **Type III**, the default in GLM, the variance estimates can be used in GLM Univariate for hypothesis testing with Type III sums of squares. Available only for the ANOVA method.

Criteria. You can specify the convergence criterion and the maximum number of iterations. Available only for the ML or REML methods.

Display. For the ANOVA method, you can choose to display sums of squares and expected mean squares. If you selected **Maximum likelihood** or **Restricted maximum likelihood**, you can display a history of the iterations.

Sum of Squares (Variance Components)

For the model, you can choose a type of sum of squares. Type III is the most commonly used and is the default.

Type I. This method is also known as the hierarchical decomposition of the sum-of-squares method. Each term is adjusted for only the term that precedes it in the model. The Type I sum-of-squares method is commonly used for:

- A balanced ANOVA model in which any main effects are specified before any first-order interaction effects, any first-order interaction effects are specified before any second-order interaction effects, and so on.
- A polynomial regression model in which any lower-order terms are specified before any higher-order terms.
- A purely nested model in which the first-specified effect is nested within the second-specified effect, the second-specified effect is nested within the third, and so on. (This form of nesting can be specified only by using syntax.)

Type III. The default. This method calculates the sums of squares of an effect in the design as the sums of squares adjusted for any other effects that do not contain it and orthogonal to any effects (if any) that contain it. The Type III sums of squares have one major advantage in that they are invariant with respect to the cell frequencies as long as the general form of estimability remains constant. Therefore, this type is often considered useful for an unbalanced model with no missing cells. In a factorial design with no missing cells, this method is equivalent to the Yates' weighted-squares-of-means technique. The Type III sum-of-squares method is commonly used for:

- Any models listed in Type I.
- Any balanced or unbalanced models with no empty cells.

Variance Components Save to New File

You can save some results of this procedure to a new IBM SPSS Statistics data file.

Variance component estimates. Saves estimates of the variance components and estimate labels to a data file or dataset. These can be used in calculating more statistics or in further analysis in the GLM procedures. For example, you can use them to calculate confidence intervals or test hypotheses.

Component covariation. Saves a variance-covariance matrix or a correlation matrix to a data file or dataset. Available only if **Maximum likelihood** or **Restricted maximum likelihood** has been specified.

Destination for created values. Allows you to specify a dataset name or external filename for the file containing the variance component estimates and/or the matrix. Datasets are available for subsequent use in the same session but are not saved as files unless explicitly saved prior to the end of the session. Dataset names must conform to variable naming rules.

You can use the MATRIX command to extract the data you need from the data file and then compute confidence intervals or perform tests.

VARCOMP Command Additional Features

The command syntax language also allows you to:

- Specify nested effects in the design (using the DESIGN subcommand).
- Include user-missing values (using the MISSING subcommand).
- Specify EPS criteria (using the CRITERIA subcommand).

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

Chapter 5. Linear Mixed Models

The Linear Mixed Models procedure expands the general linear model so that the data are permitted to exhibit correlated and nonconstant variability. The mixed linear model, therefore, provides the flexibility of modeling not only the means of the data but their variances and covariances as well.

The Linear Mixed Models procedure is also a flexible tool for fitting other models that can be formulated as mixed linear models. Such models include multilevel models, hierarchical linear models, and random coefficient models.

Example. A grocery store chain is interested in the effects of various coupons on customer spending. Taking a random sample of their regular customers, they follow the spending of each customer for 10 weeks. In each week, a different coupon is mailed to the customers. Linear Mixed Models is used to estimate the effect of different coupons on spending while adjusting for correlation due to repeated observations on each subject over the 10 weeks.

Methods. Maximum likelihood (ML) and restricted maximum likelihood (REML) estimation.

Statistics. Descriptive statistics: sample sizes, means, and standard deviations of the dependent variable and covariates for each distinct level combination of the factors. Factor-level information: sorted values of the levels of each factor and their frequencies. Also, parameter estimates and confidence intervals for fixed effects and Wald tests and confidence intervals for parameters of covariance matrices. Type I and Type III sums of squares can be used to evaluate different hypotheses. Type III is the default.

Linear Mixed Models Data Considerations

Data. The dependent variable should be quantitative. Factors should be categorical and can have numeric values or string values. Covariates and the weight variable should be quantitative. Subjects and repeated variables may be of any type.

Assumptions. The dependent variable is assumed to be linearly related to the fixed factors, random factors, and covariates. The fixed effects model the mean of the dependent variable. The random effects model the covariance structure of the dependent variable. Multiple random effects are considered independent of each other, and separate covariance matrices will be computed for each; however, model terms specified on the same random effect can be correlated. The repeated measures model the covariance structure of the residuals. The dependent variable is also assumed to come from a normal distribution.

Related procedures. Use the Explore procedure to examine the data before running an analysis. If you do not suspect there to be correlated or nonconstant variability, you can use the GLM Univariate or GLM Repeated Measures procedure. You can alternatively use the Variance Components Analysis procedure if the random effects have a variance components covariance structure and there are no repeated measures.

Obtaining a Linear Mixed Models Analysis

1. From the menus choose:

   Analyze > **Mixed Models** > **Linear...**
2. Optionally, select one or more subject variables.
3. Optionally, select one or more repeated variables.
4. Optionally, select a residual covariance structure.
5. Click **Continue**.
6. Select a dependent variable.

7. Select at least one factor or covariate.
8. Click **Fixed** or **Random** and specify at least a fixed-effects or random-effects model.

Optionally, select a weighting variable.

Linear Mixed Models Select Subjects/Repeated Variables

This dialog box allows you to select variables that define subjects and repeated observations and to choose a covariance structure for the residuals.

Subjects. A subject is an observational unit that can be considered independent of other subjects. For example, the blood pressure readings from a patient in a medical study can be considered independent of the readings from other patients. Defining subjects becomes particularly important when there are repeated measurements per subject and you want to model the correlation between these observations. For example, you might expect that blood pressure readings from a single patient during consecutive visits to the doctor are correlated.

Subjects can also be defined by the factor-level combination of multiple variables; for example, you can specify *Gender* and *Age category* as subject variables to model the belief that *males over the age of 65* are similar to each other but independent of *males under 65* and *females*.

All of the variables specified in the Subjects list are used to define subjects for the residual covariance structure. You can use some or all of the variables to define subjects for the random-effects covariance structure.

Repeated. The variables specified in this list are used to identify repeated observations. For example, a single variable *Week* might identify the 10 weeks of observations in a medical study, or *Month* and *Day* might be used together to identify daily observations over the course of a year.

Repeated Covariance type. This specifies the covariance structure for the residuals. The available structures are as follows:
- Ante-Dependence: First Order
- AR(1)
- AR(1): Heterogeneous
- ARMA(1,1)
- Compound Symmetry
- Compound Symmetry: Correlation Metric
- Compound Symmetry: Heterogeneous
- Diagonal
- Factor Analytic: First Order
- Factor Analytic: First Order, Heterogeneous
- Huynh-Feldt
- Scaled Identity
- Toeplitz
- Toeplitz: Heterogeneous
- Unstructured
- Unstructured: Correlations

See the topic Chapter 17, “Covariance Structures,” on page 99 for more information.

Linear Mixed Models Fixed Effects

Fixed Effects. There is no default model, so you must explicitly specify the fixed effects. Alternatively, you can build nested or non-nested terms.

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

Sum of Squares. The method of calculating the sums of squares. For models with no missing cells, the Type III method is most commonly used.

Build Non-Nested Terms

For the selected factors and covariates:

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

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

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

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.

Build 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 of their customers at several store locations. Since each customer frequents only one of those 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.

Sum of Squares

For the model, you can choose a type of sums of squares. Type III is the most commonly used and is the default.

Type I. This method is also known as the hierarchical decomposition of the sum-of-squares method. Each term is adjusted only for the term that precedes it in the model. Type I sums of squares are commonly used for:

- A balanced ANOVA model in which any main effects are specified before any first-order interaction effects, any first-order interaction effects are specified before any second-order interaction effects, and so on.

- A polynomial regression model in which any lower-order terms are specified before any higher-order terms.
- A purely nested model in which the first-specified effect is nested within the second-specified effect, the second-specified effect is nested within the third, and so on. (This form of nesting can be specified only by using syntax.)

Type III. The default. This method calculates the sums of squares of an effect in the design as the sums of squares adjusted for any other effects that do not contain it and orthogonal to any effects (if any) that contain it. The Type III sums of squares have one major advantage in that they are invariant with respect to the cell frequencies as long as the general form of estimability remains constant. Hence, this type of sums of squares is often considered useful for an unbalanced model with no missing cells. In a factorial design with no missing cells, this method is equivalent to the Yates' weighted-squares-of-means technique. The Type III sum-of-squares method is commonly used for:

- Any models listed in Type I.
- Any balanced or unbalanced models with no empty cells.

Linear Mixed Models Random Effects

Covariance type. This allows you to specify the covariance structure for the random-effects model. A separate covariance matrix is estimated for each random effect. The available structures are as follows:

- Ante-Dependence: First Order
- AR(1)
- AR(1): Heterogeneous
- ARMA(1,1)
- Compound Symmetry
- Compound Symmetry: Correlation Metric
- Compound Symmetry: Heterogeneous
- Diagonal
- Factor Analytic: First Order
- Factor Analytic: First Order, Heterogeneous
- Huynh-Feldt
- Scaled Identity
- Toeplitz
- Toeplitz: Heterogeneous
- Unstructured
- Unstructured: Correlation Metric
- Variance Components

See the topic Chapter 17, “Covariance Structures,” on page 99 for more information.

Random Effects. There is no default model, so you must explicitly specify the random effects. Alternatively, you can build nested or non-nested terms. You can also choose to include an intercept term in the random-effects model.

You can specify multiple random-effects models. After building the first model, click **Next** to build the next model. Click **Previous** to scroll back through existing models. Each random-effect model is assumed to be independent of every other random-effect model; that is, separate covariance matrices will be computed for each. Terms specified in the same random-effect model can be correlated.

Subject Groupings. The variables listed are those that you selected as subject variables in the Select Subjects/Repeated Variables dialog box. Choose some or all of these in order to define the subjects for the random-effects model.

Linear Mixed Models Estimation

Method. Select the maximum likelihood or restricted maximum likelihood estimation.

Iterations: The following options are available:

- **Maximum iterations.** Specify a non-negative integer.
- **Maximum step-halvings.** 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.
- **Print iteration history for every n step(s).** Displays a table containing the log-likelihood function value and parameter estimates at every *n* iteration beginning with the 0th iteration (the initial estimates). If you choose to print the iteration history, the last iteration is always printed regardless of the value of *n*.

Log-likelihood Convergence. Convergence is assumed if the absolute change or relative change in the log-likelihood function is less than the value specified, which must be non-negative. The criterion is not used if the value specified equals 0.

Parameter Convergence. Convergence is assumed if the maximum absolute change or maximum relative change in the parameter estimates is less than the value specified, which must be non-negative. The criterion is not used if the value specified equals 0.

Hessian Convergence. For the **Absolute** specification, convergence is assumed if a statistic based on the Hessian is less than the value specified. For the **Relative** specification, convergence is assumed if the statistic is less than the product of the value specified and the absolute value of the log-likelihood. The criterion is not used if the value specified equals 0.

Maximum scoring steps. Requests to use the Fisher scoring algorithm up to iteration number *n*. Specify a non-negative integer.

Singularity tolerance. This value is used as the tolerance in checking singularity. Specify a positive value.

Linear Mixed Models Statistics

Summary Statistics. Produces tables for:

- **Descriptive statistics.** Displays the sample sizes, means, and standard deviations of the dependent variable and covariates (if specified). These statistics are displayed for each distinct level combination of the factors.
- **Case Processing Summary.** Displays the sorted values of the factors, the repeated measure variables, the repeated measure subjects, and the random-effects subjects and their frequencies.

Model Statistics. Produces tables for:

- **Parameter estimates.** Displays the fixed-effects and random-effects parameter estimates and their approximate standard errors.
- **Tests for covariance parameters.** Displays the asymptotic standard errors and Wald tests for the covariance parameters.
- **Correlations of parameter estimates.** Displays the asymptotic correlation matrix of the fixed-effects parameter estimates.
- **Covariances of parameter estimates.** Displays the asymptotic covariance matrix of the fixed-effects parameter estimates.

- **Covariances of random effects.** Displays the estimated covariance matrix of random effects. This option is available only when at least one random effect is specified. If a subject variable is specified for a random effect, then the common block is displayed.
- **Covariances of residuals.** Displays the estimated residual covariance matrix. This option is available only when a repeated variable has been specified. If a subject variable is specified, the common block is displayed.
- **Contrast coefficient matrix.** This option displays the estimable functions used for testing the fixed effects and the custom hypotheses.

Confidence interval. This value is used whenever a confidence interval is constructed. Specify a value greater than or equal to 0 and less than 100. The default value is 95.

Linear Mixed Models EM Means

Estimated Marginal Means of Fitted Models. This group allows you to request model-predicted estimated marginal means of the dependent variable in the cells and their standard errors for the specified factors. Moreover, you can request that factor levels of main effects be compared.

- **Factor(s) and Factor Interactions.** This list contains factors and factor interactions that have been specified in the Fixed dialog box, plus an `OVERALL` term. Model terms built from covariates are excluded from this list.
- **Display Means for.** The procedure will compute the estimated marginal means for factors and factor interactions selected to this list. If `OVERALL` is selected, the estimated marginal means of the dependent variable are displayed, collapsing over all factors. Note that any selected factors or factor interactions remain selected unless an associated variable has been removed from the Factors list in the main dialog box.
- **Compare Main Effects.** This option allows you to request pairwise comparisons of levels of selected main effects. The Confidence Interval Adjustment allows you to apply an adjustment to the confidence intervals and significance values to account for multiple comparisons. The available methods are LSD (no adjustment), Bonferroni, and Sidak. Finally, for each factor, you can select a reference category to which comparisons are made. If no reference category is selected, all pairwise comparisons will be constructed. The options for the reference category are first, last, or custom (in which case, you enter the value of the reference category).

Linear Mixed Models Save

This dialog box allows you to save various model results to the working file.

Fixed Predicted Values. Saves variables related to the regression means without the effects.

- **Predicted values.** The regression means without the random effects.
- **Standard errors.** The standard errors of the estimates.
- **Degrees of freedom.** The degrees of freedom associated with the estimates.

Predicted Values & Residuals. Saves variables related to the model fitted value.

- **Predicted values.** The model fitted value.
- **Standard errors.** The standard errors of the estimates.
- **Degrees of freedom.** The degrees of freedom associated with the estimates.
- **Residuals.** The data value minus the predicted value.

MIXED Command Additional Features

The command syntax language also allows you to:

- Specify tests of effects versus a linear combination of effects or a value (using the `TEST` subcommand).
- Include user-missing values (using the `MISSING` subcommand).
- Compute estimated marginal means for specified values of covariates (using the `WITH` keyword of the `EMMEANS` subcommand).
- Compare simple main effects of interactions (using the `EMMEANS` subcommand).

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

Chapter 6. Generalized Linear Models

The generalized linear model expands the general linear model so that the dependent variable is linearly related to the factors and covariates via a specified link function. Moreover, the model allows for the dependent variable to have a non-normal distribution. It covers widely used statistical models, such as linear regression for normally distributed responses, logistic models for binary data, loglinear models for count data, complementary log-log models for interval-censored survival data, plus many other statistical models through its very general model formulation.

Examples. A shipping company can use generalized linear models to fit a Poisson regression to damage counts for several types of ships constructed in different time periods, and the resulting model can help determine which ship types are most prone to damage.

A car insurance company can use generalized linear models to fit a gamma regression to damage claims for cars, and the resulting model can help determine the factors that contribute the most to claim size.

Medical researchers can use generalized linear models to fit a complementary log-log regression to interval-censored survival data to predict the time to recurrence for a medical condition.

Generalized Linear Models Data Considerations

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.

Assumptions. Cases are assumed to be independent observations.

To Obtain a Generalized Linear Model

From the menus choose:

Analyze > **Generalized Linear Models** > **Generalized Linear Models...**
1. Specify a distribution and link function (see below for details on the various options).
2. On the Response tab, select a dependent variable.
3. On the Predictors tab, select factors and covariates for use in predicting the dependent variable.
4. On the Model tab, specify model effects using the selected factors and covariates.

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

Model Types

Scale Response. The following options are available:
- **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. The following options are available:
- **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. The following options are available:

- **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. The following options are available:

- **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. The following options are available:

- **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.
- **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.

- **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.

Link Functions

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 Linear Models Response

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 Linear Models Reference Category

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 Linear Models Predictors

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 target. 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 or negative binomial response with a log link if the natural log of the experience of the driver is included as an offset term.

Other combinations of distribution and link types would require other transformations of the offset variable.

Generalized Linear Models Options

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 Linear Models Model

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, such as polynomial terms involving the same covariate, 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 Linear Models Estimation

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, binomial, or multinomial distribution. The deviance and Pearson chi-square options estimate the scale parameter from the value of those statistics. Alternatively, you can specify a fixed value for the scale parameter.
- **Initial values.** The procedure will automatically compute initial values for parameters. Alternatively, you can specify initial values for the parameter estimates.
- **Covariance matrix.** The model-based estimator is the negative of the generalized inverse of the Hessian matrix. The robust (also called the Huber/White/sandwich) estimator is a "corrected" model-based estimator that provides a consistent estimate of the covariance, even when the specification of the variance and link functions is incorrect.

Iterations. The following options are available:

- **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. The following options are available

- **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 Linear Models Initial Values

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 Linear Models Statistics

Model Effects. The following options are available:

- **Analysis type.** Specify the type of analysis to produce. 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 likelihood-ratio 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 based on the assumption that parameters have an asymptotic normal distribution; profile likelihood intervals are more accurate but can be computationally expensive. The tolerance level for profile likelihood intervals is the criteria used to stop the iterative algorithm used to compute the intervals.
- **Log-likelihood function.** This controls the display format of the log-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 deviance and scaled deviance, Pearson chi-square and scaled Pearson chi-square, log-likelihood, Akaike's information criterion (AIC), finite sample corrected AIC (AICC), Bayesian information criterion (BIC), and consistent AIC (CAIC).
- **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 0th 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*.
- **Lagrange multiplier test.** Displays Lagrange multiplier test statistics for assessing the validity of a scale parameter that is computed using the deviance or Pearson chi-square, or set at a fixed number, for the normal, gamma, inverse Gaussian, and Tweedie distributions. For the negative binomial distribution, this tests the fixed ancillary parameter.

Generalized Linear Models EM Means

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 Linear Models Save

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.

- *Cook's distance*. A measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients. A large Cook's D indicates that excluding a case from computation of the regression statistics changes the coefficients substantially.
- *Leverage value*. Measures the influence of a point on the fit of the regression. The centered leverage ranges from 0 (no influence on the fit) to (N-1)/N.
- *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.
- **Standardized Pearson residual.** The Pearson residual multiplied by the square root of the inverse of the product of the scale parameter and 1−leverage for the case.
- **Deviance residual.** The square root of the contribution of a case to the Deviance statistic, with the sign of the raw residual.
- **Standardized deviance residual.** The Deviance residual multiplied by the square root of the inverse of the product of the scale parameter and 1−leverage for the case.

- **Likelihood residual.** The square root of a weighted average (based on the leverage of the case) of the squares of the standardized Pearson and standardized Deviance residuals, with the sign of the raw residual.

Generalized Linear Models Export

Export model as data. Writes a dataset in IBM SPSS Statistics format 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. You 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).
- 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 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.

Generalized Estimating Equations Data Considerations

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...**

1. 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.
2. On the Type of Model tab, specify a distribution and link function.
3. On the Response tab, select a dependent variable.
4. On the Predictors tab, select factors and covariates for use in predicting the dependent variable.
5. 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 rho for adjacent elements, rho^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. For example, if you give students standardized tests each year from 3rd through 7th grade. This structure assumes that the 3rd and 4th, 4th and 5th, 5th and 6th, and 6th and 7th grade scores will have the same correlation; 3rd and 5th, 4th and 6th, and 5th and 7th will have the same correlation; 3rd and 6th and 4th and 7th will have the same correlation. Measurements with separaration greater than *m* 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

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. The following options are available:

- **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. The following options are available:

- **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. The following options are available:

- **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. The following options are available:

- **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. The following options are available:

- **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.

- **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.
- **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.

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

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

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

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 target. 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 or negative binomial response with a log link if the natural log of the experience of the driver is included as an offset term.

Other combinations of distribution and link types would require other transformations of the offset variable.

Generalized Estimating Equations Options

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

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

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. The following options are available:

- **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. The following options are available

- **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.

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

Model Effects. The following options are available:

- **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 0th 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

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

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

Export model as data. Writes a dataset in IBM SPSS Statistics format 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. You 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. Generalized linear mixed models

Generalized linear mixed models extend the linear model so that:

- The target is linearly related to the factors and covariates via a specified link function.
- The target can have a non-normal distribution.
- The observations can be correlated.

Generalized linear mixed models cover a wide variety of models, from simple linear regression to complex multilevel models for non-normal longitudinal data.

Examples. The district school board can use a generalized linear mixed model to determine whether an experimental teaching method is effective at improving math scores. Students from the same classroom should be correlated since they are taught by the same teacher, and classrooms within the same school may also be correlated, so we can include random effects at school and class levels to account for different sources of variability. See the topic for more information.

Medical researchers can use a generalized linear mixed model to determine whether a new anticonvulsant drug can reduce a patient's rate of epileptic seizures. Repeated measurements from the same patient are typically positively correlated so a mixed model with some random effects should be appropriate. The target field, the number of seizures, takes positive integer values, so a generalized linear mixed model with a Poisson distribution and log link may be appropriate. See the topic for more information.

Executives at a cable provider of television, phone, and internet services can use a generalized linear mixed model to know more about potential customers. Since possible answers have nominal measurement levels, the company analyst uses a generalized logit mixed model with a random intercept to capture correlation between answers to the service usage questions across service types (tv, phone, internet) within a given survey responder's answers. See the topic for more information.

The Data Structure tab allows you to specify the structural relationships between records in your dataset when observations are correlated. If the records in the dataset represent independent observations, you do not need to specify anything on this tab.

Subjects. The combination of values of the specified categorical fields should uniquely define subjects within the dataset. For example, a single *Patient ID* field 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 records in the dataset.

A **subject** is an observational unit that can be considered independent of other subjects. For example, the blood pressure readings from a patient in a medical study can be considered independent of the readings from other patients. Defining subjects becomes particularly important when there are repeated measurements per subject and you want to model the correlation between these observations. For example, you might expect that blood pressure readings from a single patient during consecutive visits to the doctor are correlated.

All of the fields specified as Subjects on the Data Structure tab are used to define subjects for the residual covariance structure, and provide the list of possible fields for defining subjects for random-effects covariance structures on the Random Effect Block.

Repeated measures. The fields specified here are used to identify repeated observations. For example, a single variable *Week* might identify the 10 weeks of observations in a medical study, or *Month* and *Day* might be used together to identify daily observations over the course of a year.

Define covariance groups by. The categorical fields specified here define independent sets of repeated effects covariance parameters; one for each category defined by the cross-classification of the grouping fields. All subjects have the same covariance type; subjects within the same covariance grouping will have the same values for the parameters.

Repeated covariance type. This specifies the covariance structure for the residuals. The available structures are:

- First-order autoregressive (AR1)
- Autoregressive moving average (1,1) (ARMA11)
- Compound symmetry
- Diagonal
- Scaled identity
- Toeplitz
- Unstructured
- Variance components

Obtaining a generalized linear mixed model

This feature requires the Advanced Statistics option.

From the menus choose:

Analyze > **Mixed Models** > **Generalized Linear...**

1. Define the subject structure of your dataset on the **Data Structure** tab.
2. On the **Fields and Effects** tab, there must be a single target, which can have any measurement level, or an events/trials specification, in which case the events and trials specifications must be continuous. Optionally specify its distribution and link function, the fixed effects, and any random effects blocks, offset, or analysis weights.
3. Click **Build Options** to specify optional build settings.
4. Click **Model Options** to save scores to the active dataset and export the model to an external file.
5. Click **Run** to run the procedure and create the Model objects.

Target

These settings define the target, its distribution, and its relationship to the predictors through the link function.

Target. The target is required. It can have any measurement level, and the measurement level of the target restricts which distributions and link functions are appropriate.

- **Use number of trials as denominator.** When the target response is a number of events occurring in a set of trials, the target field contains the number of events and you can select an additional field containing the number of trials. For example, when testing a new pesticide you might expose samples of ants to different concentrations of the pesticide and then record the number of ants killed and the number of ants in each sample. In this case, the field recording the number of ants killed should be specified as the target (events) field, and the field recording the number of ants in each sample should be specified as the trials field. If the number of ants is the same for each sample, then the number of 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 record. Events should be non-negative integers, and trials should be positive integers.

- **Customize reference category.** For a categorical target, you can choose the reference category. This can affect certain output, such as parameter estimates, but it should not change the model fit. For example, if your target takes values 0, 1, and 2, by default, the procedure makes the last (highest-valued) category, or 2, the reference category. In this situation, parameter estimates should be interpreted as relating to the likelihood of category 0 or 1 *relative* to the likelihood of category 2. If you specify a custom category and your target 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 field was coded.

Target Distribution and Relationship (Link) with the Linear Model. Given the values of the predictors, the model expects the distribution of values of the target to follow the specified shape, and for the target values to be linearly related to the predictors through the specified link function. Short cuts for several common models are provided, or choose a **Custom** setting if there is a particular distribution and link function combination you want to fit that is not on the short list.

- **Linear model.** Specifies a normal distribution with an identity link, which is useful when the target can be predicted using a linear regression or ANOVA model.
- **Gamma regression.** Specifies a Gamma distribution with a log link, which should be used when the target contains all positive values and is skewed towards larger values.
- **Loglinear.** Specifies a Poisson distribution with a log link, which should be used when the target represents a count of occurrences in a fixed period of time.
- **Negative binomial regression.** Specifies a negative binomial distribution with a log link, which should be used when the target and denominator represent the number of trials required to observe *k* successes.
- **Multinomial logistic regression.** Specifies a multinomial distribution, which should be used when the target is a multi-category response. It uses either a cumulative logit link (ordinal outcomes) or a generalized logit link (multi-category nominal responses).
- **Binary logistic regression.** Specifies a binomial distribution with a logit link, which should be used when the target is a binary response predicted by a logistic regression model.
- **Binary probit.** Specifies a binomial distribution with a probit link, which should be used when the target is a binary response with an underlying normal distribution.
- **Interval censored survival.** Specifies a binomial distribution with a complementary log-log link, which is useful in survival analysis when some observations have no termination event.

Distribution

This selection specifies the distribution of the target. The ability to specify a non-normal distribution and non-identity link function is the essential improvement of the generalized linear mixed model over the linear mixed 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 a target that represents a binary response or number of events.
- **Gamma.** This distribution is appropriate for a target 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 a target 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 a target that represents a multi-category response. The form of the model will depend on the measurement level of the target.

A **nominal** target will result in a nominal multinomial model in which a separate set of model parameters are estimated for each category of the target (except the reference category). The parameter estimates for a given predictor show the relationship between that predictor and the likelihood of each category of the target, relative to the reference category.

An **ordinal** target will result in an ordinal multinomial model in which the traditional intercept term is replaced with a set of **threshold** parameters that relate to the cumulative probability of the target categories.

- **Negative binomial.** Negative binomial regression uses a negative binomial distribution with a log link, which should be used when the target represents a count of occurrences with high variance.
- **Normal.** This is appropriate for a continuous target whose values take a symmetric, bell-shaped distribution about a central (mean) value.
- **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.

Link Functions

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

- **Identity.** $f(x)=x$. The target is not transformed. This link can be used with any distribution, except the multinomial.
- **Complementary log-log.** $f(x)=\log(-\log(1-x))$. This is appropriate only with the binomial or multinomial distribution.
- **Cauchit.** $f(x) = \tan(\pi\ (x - 0.5))$. This is appropriate only with the binomial or multinomial distribution.
- **Log.** $f(x)=\log(x)$. This link can be used with any distribution, except the multinomial.
- **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 or multinomial distribution.
- **Negative log-log.** $f(x)=-\log(-\log(x))$. This is appropriate only with the binomial or multinomial 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 or multinomial 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, except the multinomial.

Fixed Effects

Fixed effects factors are generally thought of as fields whose values of interest are all represented in the dataset, and can be used for scoring. By default, fields with the predefined input role that are not specified elsewhere in the dialog are entered in the fixed effects portion of the model. Categorical (nominal, and ordinal) fields are used as factors in the model and continuous fields are used as covariates.

Enter effects into the model by selecting one or more fields in the source list and dragging to the effects list. The type of effect created depends upon which hotspot you drop the selection.

- **Main.** Dropped fields appear as separate main effects at the bottom of the effects list.
- **2-way.** All possible pairs of the dropped fields appear as 2-way interactions at the bottom of the effects list.
- **3-way.** All possible triplets of the dropped fields appear as 3-way interactions at the bottom of the effects list.
- ***.** The combination of all dropped fields appear as a single interaction at the bottom of the effects list.

Buttons to the right of the Effect Builder allow you to perform various actions.

Table 1. Effect builder button descriptions.

Icon	Description
✕	Delete terms from the fixed effects model by selecting the terms you want to delete and clicking the delete button.
↑ ↓	Reorder the terms within the fixed effects model by selecting the terms you want to reorder and clicking the up or down arrow.
✎	Add nested terms to the model using the “Add a Custom Term” dialog, by clicking on the Add a Custom Term button.

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

Add a Custom Term

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, such as polynomial terms involving the same covariate, 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.

Constructing a nested term

1. Select a factor or covariate that is nested within another factor, and then click the arrow button.
2. Click **(Within)**.
3. Select the factor within which the previous factor or covariate is nested, and then click the arrow button.
4. Click **Add Term**.

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

Random Effects

Random effects factors are fields whose values in the data file can be considered a random sample from a larger population of values. They are useful for explaining excess variability in the target. By default, if you have selected more than one subject in the Data Structure tab, a Random Effect block will be created for each subject beyond the innermost subject. For example, if you selected School, Class, and Student as subjects on the Data Structure tab, the following random effect blocks are automatically created:

- Random Effect 1: subject is school (with no effects, intercept only)
- Random Effect 2: subject is school * class (no effects, intercept only)

You can work with random effects blocks in the following ways:

1. To add a new block, click **Add Block...** This opens the "Random Effect Block" dialog.
2. To edit an existing block, select the block you want to edit and click **Edit Block...** This opens the "Random Effect Block" dialog.
3. To delete one or more blocks, select the blocks you want to delete and click the delete button.

Random Effect Block

Enter effects into the model by selecting one or more fields in the source list and dragging to the effects list. The type of effect created depends upon which hotspot you drop the selection. Categorical (nominal, and ordinal) fields are used as factors in the model and continuous fields are used as covariates.

- **Main.** Dropped fields appear as separate main effects at the bottom of the effects list.
- **2-way.** All possible pairs of the dropped fields appear as 2-way interactions at the bottom of the effects list.
- **3-way.** All possible triplets of the dropped fields appear as 3-way interactions at the bottom of the effects list.
- ***.** The combination of all dropped fields appear as a single interaction at the bottom of the effects list.

Buttons to the right of the Effect Builder allow you to perform various actions.

Table 2. Effect builder button descriptions.

Icon	Description
✕	Delete terms from the model by selecting the terms you want to delete and clicking the delete button.
↑ ↓	Reorder the terms within the model by selecting the terms you want to reorder and clicking the up or down arrow.
	Add nested terms to the model using the "Add a Custom Term" on page 61 dialog, by clicking on the Add a Custom Term button.

Include Intercept. The intercept is not included in the random effects model by default. If you can assume the data pass through the origin, you can exclude the intercept.

Define covariance groups by. The categorical fields specified here define independent sets of random effects covariance parameters; one for each category defined by the cross-classification of the grouping fields. A different set of grouping fields can be specified for each random effect block. All subjects have the same covariance type; subjects within the same covariance grouping will have the same values for the parameters.

Subject combination. This allows you to specify random effect subjects from preset combinations of subjects from the Data Structure tab. For example, if *School*, *Class*, and *Student* are defined as subjects on the Data Structure tab, and in that order, then the Subject combination dropdown list will have **None**, **School**, **School * Class**, and **School * Class * Student** as options.

Random effect covariance type. This specifies the covariance structure for the residuals. The available structures are:

- First-order autoregressive (AR1)
- Autoregressive moving average (1,1) (ARMA11)
- Compound symmetry

- Diagonal
- Scaled identity
- Toeplitz
- Unstructured
- Variance components

Weight and Offset

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

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 target. 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 or negative binomial response with a log link if the natural log of the experience of the driver is included as an offset term.

Other combinations of distribution and link types would require other transformations of the offset variable.

General Build Options

These selections specify some more advanced criteria used to build the model.

Sorting Order. These controls determine the order of the categories for the target and factors (categorical inputs) for purposes of determining the "last" category. The target sort order setting is ignored if the target is not categorical or if a custom reference category is specified on the “Target” on page 58 settings.

Stopping Rules. You can specify the maximum number of iterations the algorithm will execute. The algorithm uses a doubly iterative process that consists of an inner loop and an outer loop. The value that is specified for the maximum number of iterations applies to both loops. Specify a non-negative integer. The default is 100.

Post-Estimation Settings. These settings determine how some of the model output is computed for viewing.

- **Confidence level.** This is the level of confidence used to compute interval estimates of the model coefficients. Specify a value greater than 0 and less than 100. The default is 95.
- **Degrees of freedom.** This specifies how degrees of freedom are computed for significance tests. Choose **Fixed for all tests (Residual method)** if your sample size is sufficiently large, or the data are balanced, or the model uses a simpler covariance type; for example, scaled identity or diagonal. This is the default. Choose **Varied across tests (Satterthwaite approximation)** if your sample size is small, or the data are unbalanced, or the model uses a complicated covariance type; for example, unstructured.
- **Tests of fixed effects and coefficients.** This is the method for computing the parameter estimates covariance matrix. Choose the robust estimate if you are concerned that the model assumptions are violated.

Estimation

The model building algorithm uses a doubly iterative process that consists of an inner loop and an outer loop. The following settings apply to the inner loop.

Parameter Convergence.
Convergence is assumed if the maximum absolute change or maximum relative change in the parameter estimates is less than the value specified, which must be non-negative. The criterion is not used if the value specified equals 0.

Log-likelihood Convergence.
Convergence is assumed if the absolute change or relative change in the log-likelihood function is less than the value specified, which must be non-negative. The criterion is not used if the value specified equals 0.

Hessian Convergence.
For the **Absolute** specification, convergence is assumed if a statistic based on the Hessian is less than the value specified. For the **Relative** specification, convergence is assumed if the statistic is less than the product of the value specified and the absolute value of the log-likelihood. The criterion is not used if the value specified equals 0.

Maximum Fisher scoring steps.
Specify a non-negative integer. A value of 0 specifies the Newton-Raphson method. Values greater than 0 specify to use the Fisher scoring algorithm up to iteration number *n*, where *n* is the specified integer, and Newton-Raphson thereafter.

Singularity tolerance.
This value is used as the tolerance in checking singularity. Specify a positive value.

Note: By default, Parameter Convergence is used, where the maximum **Absolute** change at a tolerance of 1E-6 is checked. This setting might produce results that differ from the results that are obtained in versions before version 22. To reproduce results from pre-22 versions, use **Relative** for the Parameter Convergence criterion and keep the default tolerance value of 1E-6.

Estimated Means

This tab allows you to display the estimated marginal means for levels of factors and factor interactions. Estimated marginal means are not available for multinomial models.

Terms. The model terms in the Fixed Effects that are entirely comprised of categorical fields are listed here. Check each term for which you want the model to produce estimated marginal means.

- **Contrast Type.** This specifies the type of contrast to use for the levels of the contrast field. If **None** is selected, no contrasts are produced. **Pairwise** produces pairwise comparisons for all level combinations of the specified factors. This is the only available contrast for factor interactions. **Deviation** contrasts compare each level of the factor to the grand mean. **Simple** contrasts compare each level of the factor, except the last, to the last level. The "last" level is determined by the sort order for factors specified on the Build Options. Note that all of these contrast types are not orthogonal.
- **Contrast Field.** This specifies a factor, the levels of which are compared using the selected contrast type. If **None** is selected as the contrast type, no contrast field can (or need) be selected.

Continuous Fields. The listed continuous fields are extracted from the terms in the Fixed Effects that use continuous fields. When computing estimated marginal means, covariates are fixed at the specified values. Select the mean or specify a custom value.

Display estimated means in terms of. This specifies whether to compute estimated marginal means based on the original scale of the target or based on the link function transformation. **Original target scale** computes estimated marginal means for the target. Note that when the target is specified using the events/trials option, this gives the estimated marginal means for the events/trials proportion rather than for the number of events. **Link function transformation** computes estimated marginal means for the linear predictor.

Adjust for multiple comparisons using. When performing hypothesis tests with multiple contrasts, the overall significance level can be adjusted from the significance levels for the included contrasts. This 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.
- *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.
- *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.

The least significant difference method is less conservative than the sequential Sidak method, which in turn is less conservative than the sequential Bonferroni; that is, least significant difference will reject at least as many individual hypotheses as sequential Sidak, which in turn will reject at least as many individual hypotheses as sequential Bonferroni.

Save

Checked items are saved with the specified name; conflicts with existing field names are not allowed.

Predicted values. Saves the predicted value of the target. The default field name is *PredictedValue*.

Predicted probability for categorical targets. If the target is categorical, this keyword saves the predicted probabilities of the first *n* categories, up to the value specified as the **Maximum categories to save**. The calculated values are cumulative probabilities for ordinal targets. The default root name is *PredictedProbability*. To save the predicted probability of the predicted category, save the confidence (see below).

Confidence intervals. Saves upper and lower bounds of the confidence interval for the predicted value or predicted probability. For all distributions except the multinomial, this creates two variables and the default root name is *CI*, with *_Lower* and *_Upper* as the suffixes.

For the multinomial distribution and a nominal target, one field is created for each dependent variable category. This saves the lower and upper bounds of the predicted probability for the first *n* categories up to the value specified as the **Maximum categories to save**. The default root name is *CI*, and the default field names are *CI_Lower_1*, *CI_Upper_1*, *CI_Lower_2*, *CI_Upper_2*, and so on, corresponding to the order of the target categories.

For the multinomial distribution and an ordinal target, one field is created for each dependent variable category except the last (See the topic "General Build Options" on page 63 for more information.). This saves the lower and upper bounds of the cumulative predicted probability for the first *n* categories, up to but not including the last, and up to the value specified as the **Maximum categories to save**. The default root name is *CI*, and the default field names are *CI_Lower_1*, *CI_Upper_1*, *CI_Lower_2*, *CI_Upper_2*, and so on, corresponding to the order of the target categories.

Pearson residuals. Saves the Pearson residual for each record, which can be used in post-estimation diagnostics of the model fit. The default field name is *PearsonResidual*.

Confidences. Saves the confidence in the predicted value for the categorical target. The computed confidence can be based on the probability of the predicted value (the highest predicted probability) or the difference between the highest predicted probability and the second highest predicted probability. The default field name is *Confidence*.

Export model. This writes the model to an external *.zip* file. You can use this model file to apply the model information to other data files for scoring purposes. Specify a unique, valid filename. If the file specification refers to an existing file, then the file is overwritten.

Model view

The procedure creates a Model object in the Viewer. By activating (double-clicking) this object, you gain an interactive view of the model.

By default, the Model Summary view is shown. To see another model view, select it from the view thumbnails.

As an alternative to the Model object, you can generate pivot tables and charts by selecting **Pivot tables and charts** in the Output Display group on the Output tab of the Options dialog (Edit > Options). The topics that follow describe the Model object.

Model Summary

This view is a snapshot, at-a-glance summary of the model and its fit.

Table. The table identifies the target, probability distribution, and link function specified on the Target settings. If the target is defined by events and trials, the cell is split to show the events field and the trials field or fixed number of trials. Additionally the finite sample corrected Akaike information criterion (AICC) and Bayesian information criterion (BIC) are displayed.

- *Akaike Corrected*. A measure for selecting and comparing mixed models based on the -2 (Restricted) log likelihood. Smaller values indicate better models. The AICC "corrects" the AIC for small sample sizes. As the sample size increases, the AICC converges to the AIC.
- *Bayesian*. A measure for selecting and comparing models based on the -2 log likelihood. Smaller values indicate better models. The BIC also "penalizes" overparameterized models (complex models with a large number of inputs, for example), but more strictly than the AIC.

Chart. If the target is categorical, a chart displays the accuracy of the final model, which is the percentage of correct classifications.

Data Structure

This view provides a summary of the data structure you specified, and helps you to check that the subjects and repeated measures have been specified correctly. The observed information for the first subject is displayed for each subject field and repeated measures field, and the target. Additionally, the number of levels for each subject field and repeated measures field is displayed.

Predicted by Observed

For continuous targets, including targets specified as events/trials, this displays a binned scatterplot of the predicted values on the vertical axis by the observed values on the horizontal axis. Ideally, the points should lie on a 45-degree line; this view can tell you whether any records are predicted particularly badly by the model.

Classification

For categorical targets, this displays the cross-classification of observed versus predicted values in a heat map, plus the overall percent correct.

Table styles. There are several different display styles, which are accessible from the **Style** dropdown list.

- **Row percents.** This displays the row percentages (the cell counts expressed as a percent of the row totals) in the cells. This is the default.
- **Cell counts.** This displays the cell counts in the cells. The shading for the heat map is still based on the row percentages.
- **Heat map.** This displays no values in the cells, just the shading.
- **Compressed.** This displays no row or column headings, or values in the cells. It can be useful when the target has a lot of categories.

Missing. If any records have missing values on the target, they are displayed in a **(Missing)** row under all valid rows. Records with missing values do not contribute to the overall percent correct.

Multiple targets. If there are multiple categorical targets, then each target is displayed in a separate table and there is a **Target** dropdown list that controls which target to display.

Large tables. If the displayed target has more than 100 categories, no table is displayed.

Fixed Effects

This view displays the size of each fixed effect in the model.

Styles. There are different display styles, which are accessible from the **Style** dropdown list.

- **Diagram.** This is a chart in which effects are sorted from top to bottom in the order in which they were specified on the Fixed Effects settings. Connecting lines in the diagram are weighted based on effect significance, with greater line width corresponding to more significant effects (smaller *p*-values). This is the default.
- **Table.** This is an ANOVA table for the overall model and the individual model effects. The individual effects are sorted from top to bottom in the order in which they were specified on the Fixed Effects settings.

Significance. There is a Significance slider that controls which effects are shown in the view. Effects with significance values greater than the slider value are hidden. This does not change the model, but simply allows you to focus on the most important effects. By default the value is 1.00, so that no effects are filtered based on significance.

Fixed Coefficients

This view displays the value of each fixed coefficient in the model. Note that factors (categorical predictors) are indicator-coded within the model, so that **effects** containing factors will generally have multiple associated **coefficients**; one for each category except the category corresponding to the redundant coefficient.

Styles. There are different display styles, which are accessible from the **Style** dropdown list.

- **Diagram.** This is a chart which displays the intercept first, and then sorts effects from top to bottom in the order in which they were specified on the Fixed Effects settings. Within effects containing factors, coefficients are sorted by ascending order of data values. Connecting lines in the diagram are colored and weighted based on coefficient significance, with greater line width corresponding to more significant coefficients (smaller *p*-values). This is the default style.
- **Table.** This shows the values, significance tests, and confidence intervals for the individual model coefficients. After the intercept, the effects are sorted from top to bottom in the order in which they were specified on the Fixed Effects settings. Within effects containing factors, coefficients are sorted by ascending order of data values.

Multinomial. If the multinomial distribution is in effect, then the Multinomial drop-down list controls which target category to display. The sort order of the values in the list is determined by the specification on the Build Options settings.

Exponential. This displays exponential coefficient estimates and confidence intervals for certain model types, including Binary logistic regression (binomial distribution and logit link), Nominal logistic regression (multinomial distribution and logit link), Negative binomial regression (negative binomial distribution and log link), and Log-linear model (Poisson distribution and log link).

Significance. There is a Significance slider that controls which coefficients are shown in the view. Coefficients with significance values greater than the slider value are hidden. This does not change the model, but simply allows you to focus on the most important coefficients. By default the value is 1.00, so that no coefficients are filtered based on significance.

Random Effect Covariances

This view displays the random effects covariance matrix (**G**).

Styles. There are different display styles, which are accessible from the **Style** dropdown list.

- **Covariance values.** This is a heat map of the covariance matrix in which effects are sorted from top to bottom in the order in which they were specified on the Fixed Effects settings. Colors in the corrgram correspond to the cell values as shown in the key. This is the default.
- **Corrgram.** This is a heat map of the covariance matrix.
- **Compressed.** This is a heat map of the covariance matrix without the row and column headings.

Blocks. If there are multiple random effect blocks, then there is a Block dropdown list for selecting the block to display.

Groups. If a random effect block has a group specification, then there is a Group dropdown list for selecting the group level to display.

Multinomial. If the multinomial distribution is in effect, then the Multinomial drop-down list controls which target category to display. The sort order of the values in the list is determined by the specification on the Build Options settings.

Covariance Parameters

This view displays the covariance parameter estimates and related statistics for residual and random effects. These are advanced, but fundamental, results that provide information on whether the covariance structure is suitable.

Summary table. This is a quick reference for the number of parameters in the residual (**R**) and random effect (**G**) covariance matrices, the rank (number of columns) in the fixed effect (**X**) and random effect (**Z**) design matrices, and the number of subjects defined by the subject fields that define the data structure.

Covariance parameter table. For the selected effect, the estimate, standard error, and confidence interval are displayed for each covariance parameter. The number of parameters shown depends upon the covariance structure for the effect and, for random effect blocks, the number of effects in the block. If you see that the off-diagonal parameters are not significant, you may be able to use a simpler covariance structure.

Effects. If there are random effect blocks, then there is an Effect dropdown list for selecting the residual or random effect block to display. The residual effect is always available.

Groups. If a residual or random effect block has a group specification, then there is a Group dropdown list for selecting the group level to display.

Multinomial. If the multinomial distribution is in effect, then the Multinomial drop-down list controls which target category to display. The sort order of the values in the list is determined by the specification on the Build Options settings.

Estimated Means: Significant Effects

These are charts displayed for the 10 "most significant" fixed all-factor effects, starting with the three-way interactions, then the two-way interactions, and finally main effects. The chart displays the model-estimated value of the target on the vertical axis for each value of the main effect (or first listed effect in an interaction) on the horizontal axis; a separate line is produced for each value of the second listed effect in an interaction; a separate chart is produced for each value of the third listed effect in a three-way interaction; all other predictors are held constant. It provides a useful visualization of the effects of each predictor's coefficients on the target. Note that if no predictors are significant, no estimated means are produced.

Confidence. This displays upper and lower confidence limits for the marginal means, using the confidence level specified as part of the Build Options.

Estimated Means: Custom Effects

These are tables and charts for user-requested fixed all-factor effects.

Styles. There are different display styles, which are accessible from the **Style** dropdown list.

- **Diagram.** This style displays a line chart of the model-estimated value of the target on the vertical axis for each value of the main effect (or first listed effect in an interaction) on the horizontal axis; a separate line is produced for each value of the second listed effect in an interaction; a separate chart is produced for each value of the third listed effect in a three-way interaction; all other predictors are held constant.

  If contrasts were requested, another chart is displayed to compare levels of the contrast field; for interactions, a chart is displayed for each level combination of the effects other than the contrast field. For **pairwise** contrasts, it is a distance network chart; that is, a graphical representation of the comparisons table in which the distances between nodes in the network correspond to differences between samples. Yellow lines correspond to statistically significant differences; black lines correspond to non-significant differences. Hovering over a line in the network displays a tooltip with the adjusted significance of the difference between the nodes connected by the line.

  For **deviation** contrasts, a bar chart is displayed with the model-estimated value of the target on the vertical axis and the values of the contrast field on the horizontal axis; for interactions, a chart is displayed for each level combination of the effects other than the contrast field. The bars show the difference between each level of the contrast field and the overall mean, which is represented by a black horizontal line.

  For **simple** contrasts, a bar chart is displayed with the model-estimated value of the target on the vertical axis and the values of the contrast field on the horizontal axis; for interactions, a chart is displayed for each level combination of the effects other than the contrast field. The bars show the difference between each level of the contrast field (except the last) and the last level, which is represented by a black horizontal line.
- **Table.** This style displays a table of the model-estimated value of the target, its standard error, and confidence interval for each level combination of the fields in the effect; all other predictors are held constant.

  If contrasts were requested, another table is displayed with the estimate, standard error, significance test, and confidence interval for each contrast; for interactions, there a separate set of rows for each level combination of the effects other than the contrast field. Additionally, a table with the overall test results is displayed; for interactions, there is a separate overall test for each level combination of the effects other than the contrast field.

Confidence. This toggles the display of upper and lower confidence limits for the marginal means, using the confidence level specified as part of the Build Options.

Layout. This toggles the layout of the pairwise contrasts diagram. The circle layout is less revealing of contrasts than the network layout but avoids overlapping lines.

Chapter 9. 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.

Model Selection Loglinear Analysis Data Considerations

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 string variables. If a numeric variable has empty categories, use Recode to create consecutive integer values.

Obtaining a Model Selection Loglinear Analysis

From the menus choose:

Analyze > **Loglinear** > **Model Selection...**

1. Select two or more numeric categorical factors.
2. Select one or more factor variables in the Factor(s) list, and click **Define Range**.
3. Define the range of values for each factor variable.
4. Select an option in the Model Building group.

Optionally, you can select a cell weight variable to specify structural zeros.

Loglinear Analysis Define Range

You must indicate the range of categories for each factor variable. Values for Minimum and Maximum correspond to the lowest and highest categories of the factor variable. Both values must be integers, and the minimum value must be less than the maximum value. Cases with values outside of the bounds are excluded. For example, if you specify a minimum value of 1 and a maximum value of 3, only the values 1, 2, and 3 are used. Repeat this process for each factor variable.

Loglinear Analysis Model

Specify Model. A saturated model contains all factor main effects and all factor-by-factor interactions. Select **Custom** to specify a generating class for an unsaturated model.

Generating Class. A generating class is a list of the highest-order terms in which factors appear. A hierarchical model contains the terms that define the generating class and all lower-order relatives. Suppose you select variables *A*, *B*, and *C* in the Factors list and then select **Interaction** from the Build Terms drop-down list. The resulting model will contain the specified 3-way interaction *A*B*C*, the 2-way interactions *A*B*, *A*C*, and *B*C*, and main effects for *A*, *B*, and *C*. Do not specify the lower-order relatives in the generating class.

Build Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term of all selected variables. This is the default.

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

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.

Model Selection Loglinear Analysis Options

Display. You can choose **Frequencies**, **Residuals**, or both. In a saturated model, the observed and expected frequencies are equal, and the residuals are equal to 0.

Plot. For custom models, you can choose one or both types of plots, **Residuals** and **Normal Probability**. These will help determine how well a model fits the data.

Display for Saturated Model. For a saturated model, you can choose **Parameter estimates**. The parameter estimates may help determine which terms can be dropped from the model. An association table, which lists tests of partial association, is also available. This option is computationally expensive for tables with many factors.

Model Criteria. An iterative proportional-fitting algorithm is used to obtain parameter estimates. You can override one or more of the estimation criteria by specifying **Maximum iterations**, **Convergence**, or **Delta** (a value added to all cell frequencies for saturated models).

HILOGLINEAR Command Additional Features

The command syntax language also allows you to:

- Specify cell weights in matrix form (using the `CWEIGHT` subcommand).
- Generate analyses of several models with a single command (using the `DESIGN` subcommand).

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

Chapter 10. General Loglinear Analysis

The General Loglinear Analysis procedure analyzes the frequency counts of observations falling into each cross-classification category in a crosstabulation or a contingency table. Each cross-classification in the table constitutes a cell, and each categorical variable is called a factor. The dependent variable is the number of cases (frequency) in a cell of the crosstabulation, and the explanatory variables are factors and covariates. This procedure estimates maximum likelihood parameters of hierarchical and nonhierarchical loglinear models using the Newton-Raphson method. Either a Poisson or a multinomial distribution can be analyzed.

You can select up to 10 factors to define the cells of a table. A cell structure variable allows you to define structural zeros for incomplete tables, include an offset term in the model, fit a log-rate model, or implement the method of adjustment of marginal tables. Contrast variables allow computation of generalized log-odds ratios (GLOR).

Model information and goodness-of-fit statistics are automatically displayed. You can also display a variety of statistics and plots or save residuals and predicted values in the active dataset.

Example. Data from a report of automobile accidents in Florida are used to determine the relationship between wearing a seat belt and whether an injury was fatal or nonfatal. The odds ratio indicates significant evidence of a relationship.

Statistics. Observed and expected frequencies; raw, adjusted, and deviance residuals; design matrix; parameter estimates; odds ratio; log-odds ratio; GLOR; Wald statistic; and confidence intervals. Plots: adjusted residuals, deviance residuals, and normal probability.

General Loglinear Analysis Data Considerations

Data. Factors are categorical, and cell covariates are continuous. When a covariate is in the model, the mean covariate value for cases in a cell is applied to that cell. Contrast variables are continuous. They are used to compute generalized log-odds ratios. The values of the contrast variable are the coefficients for the linear combination of the logs of the expected cell counts.

A cell structure variable assigns weights. For example, if some of the cells are structural zeros, the cell structure variable has a value of either 0 or 1. Do not use a cell structure variable to weight aggregated data. Instead, choose **Weight Cases** from the Data menu.

Assumptions. Two distributions are available in General Loglinear Analysis: Poisson and multinomial.

Under the Poisson distribution assumption:

- The total sample size is not fixed before the study, or the analysis is not conditional on the total sample size.
- The event of an observation being in a cell is statistically independent of the cell counts of other cells.

Under the multinomial distribution assumption:

- The total sample size is fixed, or the analysis is conditional on the total sample size.
- The cell counts are not statistically independent.

Related procedures. Use the Crosstabs procedure to examine the crosstabulations. Use the Logit Loglinear procedure when it is natural to regard one or more categorical variables as the response variables and the others as the explanatory variables.

Obtaining a General Loglinear Analysis

1. From the menus choose:

   Analyze > **Loglinear** > **General...**
2. In the General Loglinear Analysis dialog box, select up to 10 factor variables.

Optionally, you can:

- Select cell covariates.
- Select a cell structure variable to define structural zeros or include an offset term.
- Select a contrast variable.

General Loglinear Analysis Model

Specify Model. A saturated model contains all main effects and interactions involving factor variables. It does not contain covariate terms. Select **Custom** to specify only a subset of interactions or to specify factor-by-covariate interactions.

Factors & Covariates. The factors and covariates are listed.

Terms in Model. The model depends on the nature of your data. After selecting **Custom**, you can select the main effects and interactions that are of interest in your analysis. You must indicate all of the terms to be included in the model.

Build Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term of all selected variables. This is the default.

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

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.

General Loglinear Analysis Options

The General Loglinear Analysis procedure displays model information and goodness-of-fit statistics. In addition, you can choose one or more of the following:

Display. Several statistics are available for display--observed and expected cell frequencies; raw, adjusted, and deviance residuals; a design matrix of the model; and parameter estimates for the model.

Plot. Plots, available for custom models only, include two scatterplot matrices (adjusted residuals or deviance residuals against observed and expected cell counts). You can also display normal probability and detrended normal plots of adjusted residuals or deviance residuals.

Confidence Interval. The confidence interval for parameter estimates can be adjusted.

Criteria. The Newton-Raphson method is used to obtain maximum likelihood parameter estimates. You can enter new values for the maximum number of iterations, the convergence criterion, and delta (a constant added to all cells for initial approximations). Delta remains in the cells for saturated models.

General Loglinear Analysis Save

Select the values you want to save as new variables in the active dataset. The suffix *n* in the new variable names increments to make a unique name for each saved variable.

The saved values refer to the aggregated data (cells in the contingency table), even if the data are recorded in individual observations in the Data Editor. If you save residuals or predicted values for unaggregated data, the saved value for a cell in the contingency table is entered in the Data Editor for each case in that cell. To make sense of the saved values, you should aggregate the data to obtain the cell counts.

Four types of residuals can be saved: raw, standardized, adjusted, and deviance. The predicted values can also be saved.

- *Residuals*. Also called the simple or raw residual, it is the difference between the observed cell count and its expected count.
- *Standardized residuals*. The residual divided by an estimate of its standard error. Standardized residuals are also known as Pearson residuals.
- *Adjusted residuals*. The standardized residual divided by its estimated standard error. Since the adjusted residuals are asymptotically standard normal when the selected model is correct, they are preferred over the standardized residuals for checking for normality.
- *Deviance residuals*. The signed square root of an individual contribution to the likelihood-ratio chi-square statistic (G squared), where the sign is the sign of the residual (observed count minus expected count). Deviance residuals have an asymptotic standard normal distribution.

GENLOG Command Additional Features

The command syntax language also allows you to:

- Calculate linear combinations of observed cell frequencies and expected cell frequencies and print residuals, standardized residuals, and adjusted residuals of that combination (using the `GERESID` subcommand).
- Change the default threshold value for redundancy checking (using the `CRITERIA` subcommand).
- Display the standardized residuals (using the `PRINT` subcommand).

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

Chapter 11. Logit Loglinear Analysis

The Logit Loglinear Analysis procedure analyzes the relationship between dependent (or response) variables and independent (or explanatory) variables. The dependent variables are always categorical, while the independent variables can be categorical (factors). Other independent variables (cell covariates) can be continuous, but they are not applied on a case-by-case basis. The weighted covariate mean for a cell is applied to that cell. The logarithm of the odds of the dependent variables is expressed as a linear combination of parameters. A multinomial distribution is automatically assumed; these models are sometimes called multinomial logit models. This procedure estimates parameters of logit loglinear models using the Newton-Raphson algorithm.

You can select from 1 to 10 dependent and factor variables combined. A cell structure variable allows you to define structural zeros for incomplete tables, include an offset term in the model, fit a log-rate model, or implement the method of adjustment of marginal tables. Contrast variables allow computation of generalized log-odds ratios (GLOR). The values of the contrast variable are the coefficients for the linear combination of the logs of the expected cell counts.

Model information and goodness-of-fit statistics are automatically displayed. You can also display a variety of statistics and plots or save residuals and predicted values in the active dataset.

Example. A study in Florida included 219 alligators. How does the alligators' food type vary with their size and the four lakes in which they live? The study found that the odds of a smaller alligator preferring reptiles to fish is 0.70 times lower than for larger alligators; also, the odds of selecting primarily reptiles instead of fish were highest in lake 3.

Statistics. Observed and expected frequencies; raw, adjusted, and deviance residuals; design matrix; parameter estimates; generalized log-odds ratio; Wald statistic; and confidence intervals. Plots: adjusted residuals, deviance residuals, and normal probability plots.

Logit Loglinear Analysis Data Considerations

Data. The dependent variables are categorical. Factors are categorical. Cell covariates can be continuous, but when a covariate is in the model, the mean covariate value for cases in a cell is applied to that cell. Contrast variables are continuous. They are used to compute generalized log-odds ratios (GLOR). The values of the contrast variable are the coefficients for the linear combination of the logs of the expected cell counts.

A cell structure variable assigns weights. For example, if some of the cells are structural zeros, the cell structure variable has a value of either 0 or 1. Do not use a cell structure variable to weight aggregate data. Instead, use Weight Cases on the Data menu.

Assumptions. The counts within each combination of categories of explanatory variables are assumed to have a multinomial distribution. Under the multinomial distribution assumption:

- The total sample size is fixed, or the analysis is conditional on the total sample size.
- The cell counts are not statistically independent.

Related procedures. Use the Crosstabs procedure to display the contingency tables. Use the General Loglinear Analysis procedure when you want to analyze the relationship between an observed count and a set of explanatory variables.

Obtaining a Logit Loglinear Analysis

1. From the menus choose:

Analyze > **Loglinear** > **Logit...**

2. In the Logit Loglinear Analysis dialog box, select one or more dependent variables.
3. Select one or more factor variables.

The total number of dependent and factor variables must be less than or equal to 10.

Optionally, you can:

- Select cell covariates.
- Select a cell structure variable to define structural zeros or include an offset term.
- Select one or more contrast variables.

Logit Loglinear Analysis Model

Specify Model. A saturated model contains all main effects and interactions involving factor variables. It does not contain covariate terms. Select **Custom** to specify only a subset of interactions or to specify factor-by-covariate interactions.

Factors & Covariates. The factors and covariates are listed.

Terms in Model. The model depends on the nature of your data. After selecting **Custom**, you can select the main effects and interactions that are of interest in your analysis. You must indicate all of the terms to be included in the model.

Terms are added to the design by taking all possible combinations of the dependent terms and matching each combination with each term in the model list. If **Include constant for dependent** is selected, there is also a unit term (1) added to the model list.

For example, suppose variables *D1* and *D2* are the dependent variables. A dependent terms list is created by the Logit Loglinear Analysis procedure (*D1*, *D2*, *D1*D2*). If the Terms in Model list contains *M1* and *M2* and a constant is included, the model list contains 1, *M1*, and *M2*. The resultant design includes combinations of each model term with each dependent term:

D1, *D2*, *D1*D2*

*M1*D1*, *M1*D2*, *M1*D1*D2*

*M2*D1*, *M2*D2*, *M2*D1*D2*

Include constant for dependent. Includes a constant for the dependent variable in a custom model.

Build Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term of all selected variables. This is the default.

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

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.

Logit Loglinear Analysis Options

The Logit Loglinear Analysis procedure displays model information and goodness-of-fit statistics. In addition, you can choose one or more of the following options:

Display. Several statistics are available for display: observed and expected cell frequencies; raw, adjusted, and deviance residuals; a design matrix of the model; and parameter estimates for the model.

Plot. Plots available for custom models include two scatterplot matrices (adjusted residuals or deviance residuals against observed and expected cell counts). You can also display normal probability and detrended normal plots of adjusted residuals or deviance residuals.

Confidence Interval. The confidence interval for parameter estimates can be adjusted.

Criteria. The Newton-Raphson method is used to obtain maximum likelihood parameter estimates. You can enter new values for the maximum number of iterations, the convergence criterion, and delta (a constant added to all cells for initial approximations). Delta remains in the cells for saturated models.

Logit Loglinear Analysis Save

Select the values you want to save as new variables in the active dataset. The suffix *n* in the new variable names increments to make a unique name for each saved variable.

The saved values refer to the aggregated data (to cells in the contingency table), even if the data are recorded in individual observations in the Data Editor. If you save residuals or predicted values for unaggregated data, the saved value for a cell in the contingency table is entered in the Data Editor for each case in that cell. To make sense of the saved values, you should aggregate the data to obtain the cell counts.

Four types of residuals can be saved: raw, standardized, adjusted, and deviance. The predicted values can also be saved.

- *Residuals*. Also called the simple or raw residual, it is the difference between the observed cell count and its expected count.
- *Standardized residuals*. The residual divided by an estimate of its standard error. Standardized residuals are also known as Pearson residuals.
- *Adjusted residuals*. The standardized residual divided by its estimated standard error. Since the adjusted residuals are asymptotically standard normal when the selected model is correct, they are preferred over the standardized residuals for checking for normality.
- *Deviance residuals*. The signed square root of an individual contribution to the likelihood-ratio chi-square statistic (G squared), where the sign is the sign of the residual (observed count minus expected count). Deviance residuals have an asymptotic standard normal distribution.

GENLOG Command Additional Features

The command syntax language also allows you to:

- Calculate linear combinations of observed cell frequencies and expected cell frequencies, and print residuals, standardized residuals, and adjusted residuals of that combination (using the `GERESID` subcommand).
- Change the default threshold value for redundancy checking (using the `CRITERIA` subcommand).
- Display the standardized residuals (using the `PRINT` subcommand).

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

Chapter 12. Life Tables

There are many situations in which you would want to examine the distribution of times between two events, such as length of employment (time between being hired and leaving the company). However, this kind of data usually includes some cases for which the second event isn't recorded (for example, people still working for the company at the end of the study). This can happen for several reasons: for some cases, the event simply doesn't occur before the end of the study; for other cases, we lose track of their status sometime before the end of the study; still other cases may be unable to continue for reasons unrelated to the study (such as an employee becoming ill and taking a leave of absence). Collectively, such cases are known as **censored cases**, and they make this kind of study inappropriate for traditional techniques such as *t* tests or linear regression.

A statistical technique useful for this type of data is called a follow-up **life table**. The basic idea of the life table is to subdivide the period of observation into smaller time intervals. For each interval, all people who have been observed at least that long are used to calculate the probability of a terminal event occurring in that interval. The probabilities estimated from each of the intervals are then used to estimate the overall probability of the event occurring at different time points.

Example. Is a new nicotine patch therapy better than traditional patch therapy in helping people to quit smoking? You could conduct a study using two groups of smokers, one of which received the traditional therapy and the other of which received the experimental therapy. Constructing life tables from the data would allow you to compare overall abstinence rates between the two groups to determine if the experimental treatment is an improvement over the traditional therapy. You can also plot the survival or hazard functions and compare them visually for more detailed information.

Statistics. Number entering, number leaving, number exposed to risk, number of terminal events, proportion terminating, proportion surviving, cumulative proportion surviving (and standard error), probability density (and standard error), and hazard rate (and standard error) for each time interval for each group; median survival time for each group; and Wilcoxon (Gehan) test for comparing survival distributions between groups. Plots: function plots for survival, log survival, density, hazard rate, and one minus survival.

Life Tables Data Considerations

Data. Your time variable should be quantitative. Your status variable should be dichotomous or categorical, coded as integers, with events being coded as a single value or a range of consecutive values. Factor variables should be categorical, coded as integers.

Assumptions. Probabilities for the event of interest should depend only on time after the initial event--they are assumed to be stable with respect to absolute time. That is, cases that enter the study at different times (for example, patients who begin treatment at different times) should behave similarly. There should also be no systematic differences between censored and uncensored cases. If, for example, many of the censored cases are patients with more serious conditions, your results may be biased.

Related procedures. The Life Tables procedure uses an actuarial approach to this kind of analysis (known generally as Survival Analysis). The Kaplan-Meier Survival Analysis procedure uses a slightly different method of calculating life tables that does not rely on partitioning the observation period into smaller time intervals. This method is recommended if you have a small number of observations, such that there would be only a small number of observations in each survival time interval. If you have variables that you suspect are related to survival time or variables that you want to control for (covariates), use the Cox Regression procedure. If your covariates can have different values at different points in time for the same case, use Cox Regression with Time-Dependent Covariates.

Creating Life Tables

1. From the menus choose:
   Analyze > **Survival** > **Life Tables...**
2. Select one *numeric* survival variable.
3. Specify the time intervals to be examined.
4. Select a status variable to define cases for which the terminal event has occurred.
5. Click **Define Event** to specify the value of the status variable that indicates that an event occurred.

Optionally, you can select a first-order factor variable. Actuarial tables for the survival variable are generated for each category of the factor variable.

You can also select a second-order *by factor* variable. Actuarial tables for the survival variable are generated for every combination of the first- and second-order factor variables.

Life Tables Define Events for Status Variables

Occurrences of the selected value or values for the status variable indicate that the terminal event has occurred for those cases. All other cases are considered to be censored. Enter either a single value or a range of values that identifies the event of interest.

Life Tables Define Range

Cases with values for the factor variable in the range you specify will be included in the analysis, and separate tables (and plots, if requested) will be generated for each unique value in the range.

Life Tables Options

You can control various aspects of your Life Tables analysis.

Life table(s). To suppress the display of life tables in the output, deselect **Life table(s)**.

Plot. Allows you to request plots of the survival functions. If you have defined factor variable(s), plots are generated for each subgroup defined by the factor variable(s). Available plots are survival, log survival, hazard, density, and one minus survival.

- *Survival.* Displays the cumulative survival function on a linear scale.
- *Log survival*. Displays the cumulative survival function on a logarithmic scale.
- *Hazard*. Displays the cumulative hazard function on a linear scale.
- *Density*. Displays the density function.
- *One minus survival*. Plots one-minus the survival function on a linear scale.

Compare Levels of First Factor. If you have a first-order control variable, you can select one of the alternatives in this group to perform the Wilcoxon (Gehan) test, which compares the survival of subgroups. Tests are performed on the first-order factor. If you have defined a second-order factor, tests are performed for each level of the second-order variable.

SURVIVAL Command Additional Features

The command syntax language also allows you to:

- Specify more than one dependent variable.
- Specify unequally spaced intervals.
- Specify more than one status variable.
- Specify comparisons that do not include all the factor and all the control variables.

- Calculate approximate, rather than exact, comparisons.

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

Chapter 13. Kaplan-Meier Survival Analysis

There are many situations in which you would want to examine the distribution of times between two events, such as length of employment (time between being hired and leaving the company). However, this kind of data usually includes some censored cases. Censored cases are cases for which the second event isn't recorded (for example, people still working for the company at the end of the study). The Kaplan-Meier procedure is a method of estimating time-to-event models in the presence of censored cases. The Kaplan-Meier model is based on estimating conditional probabilities at each time point when an event occurs and taking the product limit of those probabilities to estimate the survival rate at each point in time.

Example. Does a new treatment for AIDS have any therapeutic benefit in extending life? You could conduct a study using two groups of AIDS patients, one receiving traditional therapy and the other receiving the experimental treatment. Constructing a Kaplan-Meier model from the data would allow you to compare overall survival rates between the two groups to determine whether the experimental treatment is an improvement over the traditional therapy. You can also plot the survival or hazard functions and compare them visually for more detailed information.

Statistics. Survival table, including time, status, cumulative survival and standard error, cumulative events, and number remaining; and mean and median survival time, with standard error and 95% confidence interval. Plots: survival, hazard, log survival, and one minus survival.

Kaplan-Meier Data Considerations

Data. The time variable should be continuous, the status variable can be categorical or continuous, and the factor and strata variables should be categorical.

Assumptions. Probabilities for the event of interest should depend only on time after the initial event--they are assumed to be stable with respect to absolute time. That is, cases that enter the study at different times (for example, patients who begin treatment at different times) should behave similarly. There should also be no systematic differences between censored and uncensored cases. If, for example, many of the censored cases are patients with more serious conditions, your results may be biased.

Related procedures. The Kaplan-Meier procedure uses a method of calculating life tables that estimates the survival or hazard function at the time of each event. The Life Tables procedure uses an actuarial approach to survival analysis that relies on partitioning the observation period into smaller time intervals and may be useful for dealing with large samples. If you have variables that you suspect are related to survival time or variables that you want to control for (covariates), use the Cox Regression procedure. If your covariates can have different values at different points in time for the same case, use Cox Regression with Time-Dependent Covariates.

Obtaining a Kaplan-Meier Survival Analysis

1. From the menus choose:

   Analyze > **Survival** > **Kaplan-Meier...**
2. Select a time variable.
3. Select a status variable to identify cases for which the terminal event has occurred. This variable can be numeric or *short string*. Then click **Define Event.**

Optionally, you can select a factor variable to examine group differences. You can also select a strata variable, which will produce separate analyses for each level (stratum) of the variable.

Kaplan-Meier Define Event for Status Variable

Enter the value or values indicating that the terminal event has occurred. You can enter a single value, a range of values, or a list of values. The Range of Values option is available only if your status variable is numeric.

Kaplan-Meier Compare Factor Levels

You can request statistics to test the equality of the survival distributions for the different levels of the factor. Available statistics are log rank, Breslow, and Tarone-Ware. Select one of the alternatives to specify the comparisons to be made: pooled over strata, for each stratum, pairwise over strata, or pairwise for each stratum.

- *Log rank*. A test for comparing the equality of survival distributions. All time points are weighted equally in this test.
- *Breslow*. A test for comparing the equality of survival distributions. Time points are weighted by the number of cases at risk at each time point.
- *Tarone-Ware*. A test for comparing the equality of survival distributions. Time points are weighted by the square root of the number of cases at risk at each time point.
- *Pooled over strata*. Compares all factor levels in a single test to test the equality of survival curves.
- *Pairwise over strata*. Compares each distinct pair of factor levels. Pairwise trend tests are not available.
- *For each stratum*. Performs a separate test of equality of all factor levels for each stratum. If you do not have a stratification variable, the tests are not performed.
- *Pairwise for each stratum*. Compares each distinct pair of factor levels for each stratum. Pairwise trend tests are not available. If you do not have a stratification variable, the tests are not performed.

Linear trend for factor levels. Allows you to test for a linear trend across levels of the factor. This option is available only for overall (rather than pairwise) comparisons of factor levels.

Kaplan-Meier Save New Variables

You can save information from your Kaplan-Meier table as new variables, which can then be used in subsequent analyses to test hypotheses or check assumptions. You can save survival, standard error of survival, hazard, and cumulative events as new variables.

- *Survival*. Cumulative survival probability estimate. The default variable name is the prefix sur_ with a sequential number appended to it. For example, if sur_1 already exists, Kaplan-Meier assigns the variable name sur_2.
- *Standard error of survival*. Standard error of the cumulative survival estimate. The default variable name is the prefix se_ with a sequential number appended to it. For example, if se_1 already exists, Kaplan-Meier assigns the variable name se_2.
- *Hazard*. Cumulative hazard function estimate. The default variable name is the prefix haz_ with a sequential number appended to it. For example, if haz_1 already exists, Kaplan-Meier assigns the variable name haz_2.
- *Cumulative events*. Cumulative frequency of events when cases are sorted by their survival times and status codes. The default variable name is the prefix cum_ with a sequential number appended to it. For example, if cum_1 already exists, Kaplan-Meier assigns the variable name cum_2.

Kaplan-Meier Options

You can request various output types from Kaplan-Meier analysis.

Statistics. You can select statistics displayed for the survival functions computed, including survival table(s), mean and median survival, and quartiles. If you have included factor variables, separate statistics are generated for each group.

Plots. Plots allow you to examine the survival, one-minus-survival, hazard, and log-survival functions visually. If you have included factor variables, functions are plotted for each group.

- *Survival*. Displays the cumulative survival function on a linear scale.
- *One minus survival*. Plots one-minus the survival function on a linear scale.
- *Hazard*. Displays the cumulative hazard function on a linear scale.
- *Log survival*. Displays the cumulative survival function on a logarithmic scale.

KM Command Additional Features

The command syntax language also allows you to:

- Obtain frequency tables that consider cases lost to follow-up as a separate category from censored cases.
- Specify unequal spacing for the test for linear trend.
- Obtain percentiles other than quartiles for the survival time variable.

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

Chapter 14. Cox Regression Analysis

Cox Regression builds a predictive model for time-to-event data. The model produces a survival function that predicts the probability that the event of interest has occurred at a given time *t* for given values of the predictor variables. The shape of the survival function and the regression coefficients for the predictors are estimated from observed subjects; the model can then be applied to new cases that have measurements for the predictor variables. Note that information from censored subjects, that is, those that do not experience the event of interest during the time of observation, contributes usefully to the estimation of the model.

Example. Do men and women have different risks of developing lung cancer based on cigarette smoking? By constructing a Cox Regression model, with cigarette usage (cigarettes smoked per day) and gender entered as covariates, you can test hypotheses regarding the effects of gender and cigarette usage on time-to-onset for lung cancer.

Statistics. For each model: *–2LL*, the likelihood-ratio statistic, and the overall chi-square. For variables in the model: parameter estimates, standard errors, and Wald statistics. For variables not in the model: score statistics and residual chi-square.

Cox Regression Data Considerations

Data. Your time variable should be quantitative, but your status variable can be categorical or continuous. Independent variables (covariates) can be continuous or categorical; if categorical, they should be dummy- or indicator-coded (there is an option in the procedure to recode categorical variables automatically). Strata variables should be categorical, coded as integers or short strings.

Assumptions. Observations should be independent, and the hazard ratio should be constant across time; that is, the proportionality of hazards from one case to another should not vary over time. The latter assumption is known as the **proportional hazards assumption**.

Related procedures. If the proportional hazards assumption does not hold (see above), you may need to use the Cox with Time-Dependent Covariates procedure. If you have no covariates, or if you have only one categorical covariate, you can use the Life Tables or Kaplan-Meier procedure to examine survival or hazard functions for your sample(s). If you have no censored data in your sample (that is, every case experienced the terminal event), you can use the Linear Regression procedure to model the relationship between predictors and time-to-event.

Obtaining a Cox Regression Analysis

1. From the menus choose:

   Analyze > **Survival** > **Cox Regression...**
2. Select a time variable. Cases whose time values are negative are not analyzed.
3. Select a status variable, and then click **Define Event**.
4. Select one or more covariates. To include interaction terms, select all of the variables involved in the interaction and then click **>a*b>**.

Optionally, you can compute separate models for different groups by defining a strata variable.

Cox Regression Define Categorical Variables

You can specify details of how the Cox Regression procedure will handle categorical variables.

Covariates. Lists all of the covariates specified in the main dialog box, either by themselves or as part of an interaction, in any layer. If some of these are string variables or are categorical, you can use them only as categorical covariates.

Categorical Covariates. Lists variables identified as categorical. Each variable includes a notation in parentheses indicating the contrast coding to be used. String variables (denoted by the symbol < following their names) are already present in the Categorical Covariates list. Select any other categorical covariates from the Covariates list and move them into the Categorical Covariates list.

Change Contrast. Allows you to change the contrast method. Available contrast methods are:

- **Indicator.** Contrasts indicate the presence or absence of category membership. The reference category is represented in the contrast matrix as a row of zeros.
- **Simple.** Each category of the predictor variable except the reference category is compared to the reference category.
- **Difference.** Each category of the predictor variable except the first category is compared to the average effect of previous categories. Also known as reverse Helmert contrasts.
- **Helmert.** Each category of the predictor variable except the last category is compared to the average effect of subsequent categories.
- **Repeated.** Each category of the predictor variable except the first category is compared to the category that precedes it.
- **Polynomial.** Orthogonal polynomial contrasts. Categories are assumed to be equally spaced. Polynomial contrasts are available for numeric variables only.
- **Deviation.** Each category of the predictor variable except the reference category is compared to the overall effect.

If you select **Deviation**, **Simple**, or **Indicator**, select either **First** or **Last** as the reference category. Note that the method is not actually changed until you click **Change**.

String covariates must be categorical covariates. To remove a string variable from the Categorical Covariates list, you must remove all terms containing the variable from the Covariates list in the main dialog box.

Cox Regression Plots

Plots can help you to evaluate your estimated model and interpret the results. You can plot the survival, hazard, log-minus-log, and one-minus-survival functions.

- *Survival*. Displays the cumulative survival function on a linear scale.
- *Hazard*. Displays the cumulative hazard function on a linear scale.
- **Log minus log.** The cumulative survival estimate after the ln(-ln) transformation is applied to the estimate.
- *One minus survival*. Plots one-minus the survival function on a linear scale.

Because these functions depend on values of the covariates, you must use constant values for the covariates to plot the functions versus time. The default is to use the mean of each covariate as a constant value, but you can enter your own values for the plot using the Change Value control group.

You can plot a separate line for each value of a categorical covariate by moving that covariate into the Separate Lines For text box. This option is available only for categorical covariates, which are denoted by **(Cat)** after their names in the Covariate Values Plotted At list.

Cox Regression Save New Variables

You can save various results of your analysis as new variables. These variables can then be used in subsequent analyses to test hypotheses or to check assumptions.

Save Model Variables. Allows you to save the survival function and its standard error, log-minus-log estimates, hazard function, partial residuals, DfBeta(s) for the regression, and the linear predictor X*Beta as new variables.

- *Survival function*. The value of the cumulative survival function for a given time. It equals the probability of survival to that time period.
- *Log minus log survival function*. The cumulative survival estimate after the ln(-ln) transformation is applied to the estimate.
- *Hazard function*. Saves the cumulative hazard function estimate (also called the Cox-Snell residual).
- *Partial residuals*. You can plot partial residuals against survival time to test the proportional hazards assumption. One variable is saved for each covariate in the final model. Parital residuals are available only for models containing at least one covariate.
- *DfBeta(s)*. Estimated change in a coefficient if a case is removed. One variable is saved for each covariate in the final model. DfBetas are only available for models containing at least one covariate.
- *X*Beta*. Linear predictor score. The sum of the product of mean-centered covariate values and their corresponding parameter estimates for each case.

If you are running Cox with a time-dependent covariate, DfBeta(s) and the linear predictor variable X*Beta are the only variables you can save.

Export Model Information to XML File. Parameter estimates are exported to the specified file in XML format. You can use this model file to apply the model information to other data files for scoring purposes.

Cox Regression Options

You can control various aspects of your analysis and output.

Model Statistics. You can obtain statistics for your model parameters, including confidence intervals for exp(*B*) and correlation of estimates. You can request these statistics either at each step or at the last step only.

Probability for Stepwise. If you have selected a stepwise method, you can specify the probability for either entry or removal from the model. A variable is entered if the significance level of its *F*-to-enter is less than the Entry value, and a variable is removed if the significance level is greater than the Removal value. The Entry value must be less than the Removal value.

Maximum Iterations. Allows you to specify the maximum iterations for the model, which controls how long the procedure will search for a solution.

Display baseline function. Allows you to display the baseline hazard function and cumulative survival at the mean of the covariates. This display is not available if you have specified time-dependent covariates.

Cox Regression Define Event for Status Variable

Enter the value or values indicating that the terminal event has occurred. You can enter a single value, a range of values, or a list of values. The Range of Values option is available only if your status variable is numeric.

COXREG Command Additional Features

The command syntax language also allows you to:

- Obtain frequency tables that consider cases lost to follow-up as a separate category from censored cases.
- Select a reference category, other than first or last, for the deviation, simple, and indicator contrast methods.
- Specify unequal spacing of categories for the polynomial contrast method.
- Specify additional iteration criteria.
- Control the treatment of missing values.
- Specify the names for saved variables.
- Write output to an external IBM SPSS Statistics data file.
- Hold data for each split-file group in an external scratch file during processing. This can help conserve memory resources when running analyses with large datasets. This is not available with time-dependent covariates.

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

Chapter 15. Computing Time-Dependent Covariates

There are certain situations in which you would want to compute a Cox Regression model but the proportional hazards assumption does not hold. That is, hazard ratios change across time; the values of one (or more) of your covariates are different at different time points. In such cases, you need to use an extended Cox Regression model, which allows you to specify **time-dependent covariates**.

In order to analyze such a model, you must first define your time-dependent covariate. (Multiple time-dependent covariates can be specified using command syntax.) To facilitate this, a system variable representing time is available. This variable is called *T_*. You can use this variable to define time-dependent covariates in two general ways:

- If you want to test the proportional hazards assumption with respect to a particular covariate or estimate an extended Cox regression model that allows nonproportional hazards, you can do so by defining your time-dependent covariate as a function of the time variable *T_* and the covariate in question. A common example would be the simple product of the time variable and the covariate, but more complex functions can be specified as well. Testing the significance of the coefficient of the time-dependent covariate will tell you whether the proportional hazards assumption is reasonable.
- Some variables may have different values at different time periods but aren't systematically related to time. In such cases, you need to define a **segmented time-dependent covariate**, which can be done using **logical expressions**. Logical expressions take the value 1 if true and 0 if false. Using a series of logical expressions, you can create your time-dependent covariate from a set of measurements. For example, if you have blood pressure measured once a week for the four weeks of your study (identified as *BP1* to *BP4*), you can define your time-dependent covariate as (*T_* < 1) * *BP1* + (*T_* >= 1 & *T_* < 2) * *BP2* + (*T_* >= 2 & *T_* < 3) * *BP3* + (*T_* >= 3 & *T_* < 4) * *BP4*. Notice that exactly one of the terms in parentheses will be equal to 1 for any given case and the rest will all equal 0. In other words, this function means that if time is less than one week, use *BP1*; if it is more than one week but less than two weeks, use *BP2*, and so on.

In the Compute Time-Dependent Covariate dialog box, you can use the function-building controls to build the expression for the time-dependent covariate, or you can enter it directly in the Expression for T_COV_ text area. Note that string constants must be enclosed in quotation marks or apostrophes, and numeric constants must be typed in American format, with the dot as the decimal delimiter. The resulting variable is called *T_COV_* and should be included as a covariate in your Cox Regression model.

Computing a Time-Dependent Covariate

1. From the menus choose:

   Analyze > **Survival** > **Cox w/ Time-Dep Cov...**
2. Enter an expression for the time-dependent covariate.
3. Click **Model** to proceed with your Cox Regression.

Note: Be sure to include the new variable *T_COV_* as a covariate in your Cox Regression model.

See the topic Chapter 14, “Cox Regression Analysis,” on page 89 for more information.

Cox Regression with Time-Dependent Covariates Additional Features

The command syntax language also allows you to specify multiple time-dependent covariates. Other command syntax features are available for Cox Regression with or without time-dependent covariates.

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

Chapter 16. Categorical Variable Coding Schemes

In many procedures, you can request automatic replacement of a categorical independent variable with a set of contrast variables, which will then be entered or removed from an equation as a block. You can specify how the set of contrast variables is to be coded, usually on the CONTRAST subcommand. This appendix explains and illustrates how different contrast types requested on CONTRAST actually work.

Deviation

Deviation from the grand mean. In matrix terms, these contrasts have the form:

```
   mean (  1/k    1/k   ...   1/k   1/k)
  df(1) (1-1/k   -1/k   ...  -1/k  -1/k)
  df(2) ( -1/k  1-1/k   ...  -1/k  -1/k)
      .           .
      .           .
df(k-1) ( -1/k   -1/k   ... 1-1/k  -1/k)
```

where *k* is the number of categories for the independent variable and the last category is omitted by default. For example, the deviation contrasts for an independent variable with three categories are as follows:

```
( 1/3   1/3   1/3)
( 2/3  -1/3  -1/3)
(-1/3   2/3  -1/3)
```

To omit a category other than the last, specify the number of the omitted category in parentheses after the DEVIATION keyword. For example, the following subcommand obtains the deviations for the first and third categories and omits the second:

```
/CONTRAST(FACTOR)=DEVIATION(2)
```

Suppose that *factor* has three categories. The resulting contrast matrix will be

```
( 1/3   1/3   1/3)
( 2/3  -1/3  -1/3)
(-1/3  -1/3   2/3)
```

Simple

Simple contrasts. Compares each level of a factor to the last. The general matrix form is

```
   mean (1/k  1/k   ...   1/k  1/k)
  df(1) (  1    0   ...     0   -1)
  df(2) (  0    1   ...     0   -1)
      .       .
      .       .
df(k-1) (  0    0   ...     1   -1)
```

where *k* is the number of categories for the independent variable. For example, the simple contrasts for an independent variable with four categories are as follows:

```
(1/4   1/4   1/4  1/4)
(  1     0     0   -1)
(  0     1     0   -1)
(  0     0     1   -1)
```

To use another category instead of the last as a reference category, specify in parentheses after the SIMPLE keyword the sequence number of the reference category, which is not necessarily the value associated with that category. For example, the following CONTRAST subcommand obtains a contrast matrix that omits the second category:

```
/CONTRAST(FACTOR) = SIMPLE(2)
```

Suppose that *factor* has four categories. The resulting contrast matrix will be

```
(1/4   1/4   1/4  1/4)
(  1    -1     0    0)
(  0    -1     1    0)
(  0    -1     0    1)
```

Helmert

Helmert contrasts. Compares categories of an independent variable with the mean of the subsequent categories. The general matrix form is

```
   mean (1/k       1/k   ...       1/k       1/k       1/k)
  df(1) (  1  -1/(k-1)   ...  -1/(k-1)  -1/(k-1)  -1/(k-1))
  df(2) (  0         1   ...  -1/(k-2)  -1/(k-2)  -1/(k-2))
      .              .
      .              .
df(k-2) (  0         0   ...         1      -1/2      -1/2)
df(k-1) (  0         0   ...         0         1        -1)
```

where *k* is the number of categories of the independent variable. For example, an independent variable with four categories has a Helmert contrast matrix of the following form:

```
(1/4   1/4   1/4   1/4)
(  1  -1/3  -1/3  -1/3)
(  0     1  -1/2  -1/2)
(  0     0     1    -1)
```

Difference

Difference or reverse Helmert contrasts. Compares categories of an independent variable with the mean of the previous categories of the variable. The general matrix form is

```
   mean (      1/k       1/k       1/k  ...  1/k)
  df(1) (       -1         1         0  ...    0)
  df(2) (     -1/2      -1/2         1  ...    0)
      .                    .
      .                    .
df(k-1) (-1/(k-1)  -1/(k-1)  -1/(k-1)  ...    1)
```

where *k* is the number of categories for the independent variable. For example, the difference contrasts for an independent variable with four categories are as follows:

```
( 1/4   1/4   1/4   1/4)
(  -1     1     0     0)
(-1/2  -1/2     1     0)
(-1/3  -1/3  -1/3     1)
```

Polynomial

Orthogonal polynomial contrasts. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; the third degree of freedom, the cubic; and so on, for the higher-order effects.

You can specify the spacing between levels of the treatment measured by the given categorical variable. Equal spacing, which is the default if you omit the metric, can be specified as consecutive integers from 1 to *k*, where *k* is the number of categories. If the variable *drug* has three categories, the subcommand

```
/CONTRAST(DRUG)=POLYNOMIAL
```

is the same as

```
/CONTRAST(DRUG)=POLYNOMIAL(1,2,3)
```

Equal spacing is not always necessary, however. For example, suppose that *drug* represents different dosages of a drug given to three groups. If the dosage administered to the second group is twice that given to the first group and the dosage administered to the third group is three times that given to the first group, the treatment categories are equally spaced, and an appropriate metric for this situation consists of consecutive integers:

```
/CONTRAST(DRUG)=POLYNOMIAL(1,2,3)
```

If, however, the dosage administered to the second group is four times that given to the first group, and the dosage administered to the third group is seven times that given to the first group, an appropriate metric is

```
/CONTRAST(DRUG)=POLYNOMIAL(1,4,7)
```

In either case, the result of the contrast specification is that the first degree of freedom for *drug* contains the linear effect of the dosage levels and the second degree of freedom contains the quadratic effect.

Polynomial contrasts are especially useful in tests of trends and for investigating the nature of response surfaces. You can also use polynomial contrasts to perform nonlinear curve fitting, such as curvilinear regression.

Repeated

Compares adjacent levels of an independent variable. The general matrix form is

```
   mean (1/k  1/k  1/k  ...  1/k  1/k)
  df(1) (  1   -1    0  ...    0    0)
  df(2) (  0    1   -1  ...    0    0)
      .        .
      .        .
df(k-1) (  0    0    0  ...    1   -1)
```

where *k* is the number of categories for the independent variable. For example, the repeated contrasts for an independent variable with four categories are as follows:

```
(1/4   1/4   1/4  1/4)
(  1    -1     0    0)
(  0     1    -1    0)
(  0     0     1   -1)
```

These contrasts are useful in profile analysis and wherever difference scores are needed.

Special

A user-defined contrast. Allows entry of special contrasts in the form of square matrices with as many rows and columns as there are categories of the given independent variable. For MANOVA and LOGLINEAR, the first row entered is always the mean, or constant, effect and represents the set of weights indicating how to average other independent variables, if any, over the given variable. Generally, this contrast is a vector of ones.

The remaining rows of the matrix contain the special contrasts indicating the comparisons between categories of the variable. Usually, orthogonal contrasts are the most useful. Orthogonal contrasts are statistically independent and are nonredundant. Contrasts are orthogonal if:

- For each row, contrast coefficients sum to 0.
- The products of corresponding coefficients for all pairs of disjoint rows also sum to 0.

For example, suppose that treatment has four levels and that you want to compare the various levels of treatment with each other. An appropriate special contrast is

```
(1   1   1   1)   weights for mean calculation
(3  -1  -1  -1)   compare 1st with 2nd through 4th
(0   2  -1  -1)   compare 2nd with 3rd and 4th
(0   0   1  -1)   compare 3rd with 4th
```

which you specify by means of the following CONTRAST subcommand for MANOVA, LOGISTIC REGRESSION, and COXREG:

```
/CONTRAST(TREATMNT)=SPECIAL( 1  1  1  1
                             3 -1 -1 -1
                             0  2 -1 -1
                             0  0  1 -1 )
```

For LOGLINEAR, you need to specify:

```
/CONTRAST(TREATMNT)=BASIS SPECIAL( 1  1  1  1
                                   3 -1 -1 -1
                                   0  2 -1 -1
                                   0  0  1 -1 )
```

Each row except the means row sums to 0. Products of each pair of disjoint rows sum to 0 as well:

```
Rows 2 and 3:  (3)(0) + (–1)(2) + (–1)(–1) + (–1)(–1) = 0
Rows 2 and 4:  (3)(0) + (–1)(0) + (–1)(1) + (–1)(–1) = 0
Rows 3 and 4:  (0)(0) + (2)(0) + (–1)(1) + (–1)(–1) = 0
```

The special contrasts need not be orthogonal. However, they must not be linear combinations of each other. If they are, the procedure reports the linear dependency and ceases processing. Helmert, difference, and polynomial contrasts are all orthogonal contrasts.

Indicator

Indicator variable coding. Also known as dummy coding, this is not available in LOGLINEAR or MANOVA. The number of new variables coded is *k*–1. Cases in the reference category are coded 0 for all *k*–1 variables. A case in the *i* th category is coded 0 for all indicator variables except the *i* th, which is coded 1.

Chapter 17. Covariance Structures

This section provides additional information on covariance structures.

Ante-Dependence: First-Order. This covariance structure has heterogenous variances and heterogenous correlations between adjacent elements. The correlation between two nonadjacent elements is the product of the correlations between the elements that lie between the elements of interest.

$$\begin{array}{llll}
(\sigma_1^{2} & \sigma_2\sigma_1\rho_1 & \sigma_3\sigma_1\rho_1\rho_2 & \sigma_4\sigma_1\rho_1\rho_2\rho_3) \\
(\sigma_2\sigma_1\rho_1 & \sigma_2^{2} & \sigma_3\sigma_2\rho_2 & \sigma_4\sigma_2\rho_2\rho_3) \\
(\sigma_3\sigma_1\rho_1\rho_2 & \sigma_3\sigma_2\rho_2 & \sigma_3^{2} & \sigma_4\sigma_3\rho_3) \\
(\sigma_4\sigma_1\rho_1\rho_2\rho_3 & \sigma_4\sigma_2\rho_2\rho_3 & \sigma_4\sigma_3\rho_3 & \sigma_4^{2})
\end{array}$$

AR(1). This is a first-order autoregressive structure with homogenous variances. The correlation between any two elements is equal to rho for adjacent elements, rho^2 for elements that are separated by a third, and so on. is constrained so that –1<<1.

$$\begin{array}{llll}
(\sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3) \\
(\sigma^2\rho & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2) \\
(\sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \sigma^2\rho) \\
(\sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2)
\end{array}$$

AR(1): Heterogenous. This is a first-order autoregressive structure with heterogenous variances. The correlation between any two elements is equal to r for adjacent elements, r^2 for two elements separated by a third, and so on. is constrained to lie between –1 and 1.

$$\begin{array}{llll}
(\sigma_1^{2} & \sigma_2\sigma_1\rho & \sigma_3\sigma_1\rho^2 & \sigma_4\sigma_1\rho^3) \\
(\sigma_2\sigma_1\rho & \sigma_2^{2} & \sigma_3\sigma_2\rho & \sigma_4\sigma_2\rho^2) \\
(\sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^{2} & \sigma_4\sigma_3\rho) \\
(\sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^{2})
\end{array}$$

ARMA(1,1). This is a first-order autoregressive moving average structure. It has homogenous variances. The correlation between two elements is equal to * for adjacent elements, *(^{2}) for elements separated by a third, and so on. and are the autoregressive and moving average parameters, respectively, and their values are constrained to lie between –1 and 1, inclusive.

$$\begin{array}{llll}
(\sigma^2 & \sigma^2\phi\rho & \sigma^2\phi\rho^2 & \sigma^2\phi\rho^3) \\
(\sigma^2\phi\rho & \sigma^2 & \sigma^2\phi\rho & \sigma^2\phi\rho^2) \\
(\sigma^2\phi\rho^2 & \sigma^2\phi\rho & \sigma^2 & \sigma^2\phi\rho) \\
(\sigma^2\phi\rho^3 & \sigma^2\phi\rho^2 & \sigma^2\phi\rho & \sigma^2)
\end{array}$$

Compound Symmetry. This structure has constant variance and constant covariance.

$$\begin{array}{llll}
(\sigma^2 + \sigma_1^{2} & \sigma_1 & \sigma_1 & \sigma_1) \\
(\sigma_1 & \sigma^2 + \sigma_1^{2} & \sigma_1 & \sigma_1) \\
(\sigma_1 & \sigma_1 & \sigma^2 + \sigma_1^{2} & \sigma_1) \\
(\sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1^{2})
\end{array}$$

Compound Symmetry: Correlation Metric. This covariance structure has homogenous variances and homogenous correlations between elements.

(σ^2	$\sigma^2\rho$	$\sigma^2\rho$	$\sigma^2\rho$)
($\sigma^2\rho$	σ^2	$\sigma^2\rho$	$\sigma^2\rho$)
($\sigma^2\rho$	$\sigma^2\rho$	σ^2	$\sigma^2\rho$)
($\sigma^2\rho$	$\sigma^2\rho$	$\sigma^2\rho$	σ^2)

Compound Symmetry: Heterogenous. This covariance structure has heterogenous variances and constant correlation between elements.

(σ_1^2	$\sigma_2\sigma_1\rho$	$\sigma_3\sigma_1\rho$	$\sigma_4\sigma_1\rho$)
($\sigma_2\sigma_1\rho$	σ_2^2	$\sigma_3\sigma_2\rho$	$\sigma_4\sigma_2\rho$)
($\sigma_3\sigma_1\rho$	$\sigma_3\sigma_2\rho$	σ_3^2	$\sigma_4\sigma_3\rho$)
($\sigma_4\sigma_1\rho$	$\sigma_4\sigma_2\rho$	$\sigma_4\sigma_3\rho$	σ_4^2)

Diagonal. This covariance structure has heterogenous variances and zero correlation between elements.

(σ_1^2	0	0	0)
(0	σ_2^2	0	0)
(0	0	σ_3^2	0)
(0	0	0	σ_4^2)

Factor Analytic: First-Order. This covariance structure has heterogenous variances that are composed of a term that is heterogenous across elements and a term that is homogenous across elements. The covariance between any two elements is the square root of the product of their heterogenous variance terms.

($\lambda_1^2 + d$	$\lambda_2\lambda_1$	$\lambda_3\lambda_1$	$\lambda_4\lambda_1$)
($\lambda_2\lambda_1$	$\lambda_2^2 + d$	$\lambda_3\lambda_2$	$\lambda_4\lambda_2$)
($\lambda_3\lambda_1$	$\lambda_3\lambda_2$	$\lambda_3^2 + d$	$\lambda_4\lambda_3$)
($\lambda_4\lambda_1$	$\lambda_4\lambda_2$	$\lambda_4\lambda_3$	$\lambda_4^2 + d$)

Factor Analytic: First-Order, Heterogenous. This covariance structure has heterogenous variances that are composed of two terms that are heterogenous across elements. The covariance between any two elements is the square root of the product of the first of their heterogenous variance terms.

($\lambda_1^2 + d_1$	$\lambda_2\lambda_1$	$\lambda_3\lambda_1$	$\lambda_4\lambda_1$)
($\lambda_2\lambda_1$	$\lambda_2^2 + d_2$	$\lambda_3\lambda_2$	$\lambda_4\lambda_2$)
($\lambda_3\lambda_1$	$\lambda_3\lambda_2$	$\lambda_3^2 + d_3$	$\lambda_4\lambda_3$)
($\lambda_4\lambda_1$	$\lambda_4\lambda_2$	$\lambda_4\lambda_3$	$\lambda_4^2 + d_4$)

Huynh-Feldt. This is a "circular" matrix in which the covariance between any two elements is equal to the average of their variances minus a constant. Neither the variances nor the covariances are constant.

(σ_1^2	$[\sigma_1^2 + \sigma_2^2]/2 - \lambda$	$[\sigma_1^2 + \sigma_3^2]/2 - \lambda$	$[\sigma_1^2 + \sigma_4^2]/2 - \lambda$)
($[\sigma_1^2 + \sigma_2^2]/2 - \lambda$	σ_2^2	$[\sigma_2^2 + \sigma_3^2]/2 - \lambda$	$[\sigma_2^2 + \sigma_4^2]/2 - \lambda$)
($[\sigma_1^2 + \sigma_3^2]/2 - \lambda$	$[\sigma_2^2 + \sigma_3^2]/2 - \lambda$	σ_3^2	$[\sigma_3^2 + \sigma_4^2]/2 - \lambda$)
($[\sigma_1^2 + \sigma_4^2]/2 - \lambda$	$[\sigma_2^2 + \sigma_4^2]/2 - \lambda$	$[\sigma_3^2 + \sigma_4^2]/2 - \lambda$	σ_4^2)

Scaled Identity. This structure has constant variance. There is assumed to be no correlation between any elements.

(σ^2	0	0	0)
(0	σ^2	0	0)
(0	0	σ^2	0)

(0	0	0	σ^2)

Toeplitz. This covariance structure has homogenous variances and heterogenous correlations between elements. The correlation between adjacent elements is homogenous across pairs of adjacent elements. The correlation between elements separated by a third is again homogenous, and so on.

(σ^2	$\sigma^2\rho_1$	$\sigma^2\rho_2$	$\sigma^2\rho_3$)
($\sigma^2\rho_1$	σ^2	$\sigma^2\rho_1$	$\sigma^2\rho_2$)
($\sigma^2\rho_2$	$\sigma^2\rho_1$	σ^2	$\sigma^2\rho_1$)
($\sigma^2\rho_3$	$\sigma^2\rho_2$	$\sigma^2\rho_1$	σ^2)

Toeplitz: Heterogenous. This covariance structure has heterogenous variances and heterogenous correlations between elements. The correlation between adjacent elements is homogenous across pairs of adjacent elements. The correlation between elements separated by a third is again homogenous, and so on.

($\sigma_1{}^2$	$\sigma_2\sigma_1\rho_1$	$\sigma_3\sigma_1\rho_2$	$\sigma_4\sigma_1\rho_3$)
($\sigma_2\sigma_1\rho_1$	$\sigma_2{}^2$	$\sigma_3\sigma_2\rho_1$	$\sigma_4\sigma_2\rho_2$)
($\sigma_3\sigma_1\rho_2$	$\sigma_3\sigma_2\rho_1$	$\sigma_3{}^2$	$\sigma_4\sigma_3\rho_1$)
($\sigma_4\sigma_1\rho_3$	$\sigma_4\sigma_2\rho_2$	$\sigma_4\sigma_3\rho_1$	$\sigma_4{}^2$)

Unstructured. This is a completely general covariance matrix.

($\sigma_1{}^2$	$\sigma_{2\ 1}$	σ_{31}	σ_{41})
($\sigma_{2\ 1}$	$\sigma_2{}^2$	σ_{32}	$\sigma_{4\ 2}$)
(σ_{31}	σ_{32}	$\sigma_3{}^2$	$\sigma_{4\ 3}$)
(σ_{41}	$\sigma_{4\ 2}$	$\sigma_{4\ 3}$	$\sigma_4{}^2$)

Unstructured: Correlation Metric. This covariance structure has heterogenous variances and heterogenous correlations.

($\sigma_1{}^2$	$\sigma_2\sigma_1\rho_{21}$	$\sigma_3\sigma_1\rho_{31}$	$\sigma_4\sigma_1\rho_{41}$)
($\sigma_2\sigma_1\rho_{21}$	$\sigma_2{}^2$	$\sigma_3\sigma_2\rho_{32}$	$\sigma_4\sigma_2\rho_{42}$)
($\sigma_3\sigma_1\rho_{31}$	$\sigma_3\sigma_2\rho_{32}$	$\sigma_3{}^2$	$\sigma_4\sigma_3\rho_{43}$)
($\sigma_4\sigma_1\rho_{41}$	$\sigma_4\sigma_2\rho_{42}$	$\sigma_4\sigma_3\rho_{43}$	$\sigma_4{}^2$)

Variance Components. This structure assigns a scaled identity (ID) structure to each of the specified random effects.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who want to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs.

© Copyright IBM Corp. _enter the year or years_. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Index

Printed in USA