IBM SPSS Advanced Statistics

More accurately analyze complex relationships

Business benefits

- Move beyond basic analysis

- Build flexible models using a wealth of model-building options

- Achieve more accurate predictive models using a wide range of modeling techniques

- Identify random effects

- Analyze results using various methods

Make your analysis more accurate and reach more dependable conclusions with statistics designed to fit the inherent characteristics of data describing complex relationships. IBM SPSS Advanced Statistics* provides a powerful set of sophisticated univariate and multivariate analytical techniques for real-world problems, such as:

- **Medical research** – Analyze patient survival rates
- **Manufacturing** – Assess production processes
- **Pharmaceutical** – Report test results to the FDA
- **Market research** – Determine product interest levels

Access a wide range of powerful models

In addition to general linear models (GLM), mixed models procedures, generalized linear models (GENLIN) and generalized estimating equations (GEE) procedures, IBM SPSS Advanced Statistics now offers generalized linear mixed models (GLMM).

Generalized linear mixed models include a wide variety of models, from simple linear regression to complex multilevel models for non-normal longitudinal data. The GLMM procedure produces more accurate models when predicting nonlinear outcomes (for example, what product a customer is likely to buy) by taking into account hierarchical data structures (for example customer nested within an organization).

GENLIN includes widely used statistical models, such as linear regression for normally distributed responses, logistic models for binary data, and loglinear models for count data. This procedure also offers many useful statistical models through its very general model formulation, such as ordinal regression, Tweedie regression, Poisson regression, Gamma regression, and negative binomial regression. GEE procedures extend generalized linear models to accommodate correlated longitudinal data and clustered data.

GENLIN and GEE provide a common framework for the following outcomes:

- Numerical: Linear regression, analysis of variance, analysis of covariance, repeated measures analysis, and Gamma regression
- Count data: Loglinear models, logistic regression, probit regression, Poisson regression, and negative binomial regression
- Ordinal data: Ordinal regression
- Event/trial data: Logistic regression
- Claim data: Inverse Gaussian regression
- Combination of discrete and continuous outcomes: Tweedie regression
- Correlated responses within subjects: GEE or correlated response models

Get more accurate predictive models when working with nested-structure data

The linear mixed models procedure expands upon the models used in the GLM procedure so that you can analyze data that exhibit correlation and non-constant variability. This procedure enables you to model not only means but also variances and covariances in your data.

The procedure's flexibility allows you to formulate a wide variety of models, including fixed effects ANOVA models, randomized complete blocks designs, split-plot designs, purely random effects models, random coefficient models, multilevel analyses, unconditional linear growth models, linear growth models with person-level covariates, repeated measures analyses, and repeated measures analyses with time-dependent covariates. Work with repeated measures designs, including incomplete repeated measurements in which the number of observations varies across subjects.

Build flexible models

The GLM procedure enables you to describe the relationship between a dependent variable and a set of independent variables. Models include linear regression, ANOVA, ANCOVA, MANOVA, and MANCOVA. GLM also includes capabilities for repeated measures, mixed models, post hoc tests and post hoc tests for repeated measures, four types of sums of squares, and pairwise comparisons of expected marginal means,

as well as the sophisticated handling of missing cells, and the option to save design matrices and effect files.

IBM SPSS Advanced Statistics is available for installation as client-only software but, for greater performance and scalability, a server-based version is also available.

Apply more sophisticated models

Use IBM SPSS Advanced Statistics when your data do not conform to the assumptions required by simpler techniques. IBM SPSS Advanced Statistics has loglinear and hierarchical loglinear analysis for modeling multiway tables of count data. The general loglinear analysis procedure helps you analyze the frequency counts of observations falling into each cross-classification category in a crosstabulation or contingency table. You can select up to 10 factors to define the cells of a table. Model information and goodness-of-fit statistics are shown automatically. Display a variety of statistics and plots, or save residuals and predicted values in the working data file.

Analyze event history and duration data

You can examine lifetime or duration data to understand terminal events, such as part failure, death, or survival. IBM SPSS Advanced Statistics includes Kaplan-Meier and Cox regression, state-of-the-art survival procedures. Use Kaplan-Meier estimations to gauge the length of time to an event; use Cox regression to perform proportional hazard regression with time-to-response or duration response as the dependent variable. These procedures, along with life tables analyses, provide a flexible and comprehensive set of techniques for working with your survival data.

Gain greater value with collaboration

To share and re-use assets efficiently, protect them in ways that meet internal and external compliance requirements, and publish results so that a greater number of business users can view and interact with them, consider augmenting your IBM SPSS Statistics software with IBM® SPSS® Collaboration and Deployment Services (formerly PASW® Collaboration and Deployment Services). More information about these valuable capabilities can be by downloading the "Collaboration" brochure at www.spss.com/software/deployment/cds.

Features
Generalized Linear Mixed Models (GLMM)

GLMM extends the linear model so that: 1) the target is linearly related to the factors and covariates through a specified link function, 2) the target can have a non-normal distribution and 3) the observations can be correlated.

- Specify the subject structure for repeated measurements and how the errors of the repeated measurements are correlated
- Choose among the 8 covariance types
- Specify the target, optional offset and optional analysis (regression) weight
- Choose among the following probability distributions: binomial, gamma, inverse Gaussian, multinomial, negative binomial, normal, Poisson
- Choose among the following link functions: identity, complementary log-log. log-link, log complement, logit, negative log-log, power, probit
- Specify (optional) fixed model effects, including the intercept
- Specify the random effects in the mixed model
- Display estimated marginal means of the target for all level combinations of a set of factors
- Save a file containing the scoring model
- Write optional temporary fields to the active dataset

GENLIN and GEE

GENLIN procedures provide a unifying framework that includes classical linear models with normally distributed dependent variables, logistic and probit models for binary data and loglinear models for count data, as well as various other nonstandard regression-type models. GEE procedures extend the generalized linear model to correlated longitudinal data and clustered data. More particularly, GEE procedures model correlations within subjects.

- Users benefit from having a common framework for the following outcomes:
  - Continuous outcomes: Linear regression, analysis of variance, analysis of covariance, repeated measures analysis and Gamma regression
  - Ordinal data: Ordinal regression
  - Count data: Loglinear models, logistic regression, probit regression, Poisson regression and negative binomial regression
  - Event/trial data: Logistic regression
  - Claim data: Inverse Gaussian regression
  - Combination of discrete and continuous outcomes: Tweedie regression
  - Correlated responses within subjects: GEE or correlated response models

- The MODEL subcommand is used to specify model effects, an offset or scale weight variable if either exists, the probability distribution and the link function
  - Offers an option to include or exclude the intercept
  - Specifies an offset variable or fixes the offset at a number
  - Specifies a variable that contains Omega weight values for the scale parameter
  - Enables users to choose from the following probability distributions: Binomial, Gamma, inverse Gaussian, negative binomial, normal, multinomial ordinal, Tweedie and Poisson
  - Offers the following link functions: Complementary log-log, identity, log, log complement, logit, negative binomial, negative log-log, odds power, probit, cumulative logit and power

- The CRITERIA subcommand controls statistical criteria for GENLIN and specifies numerical tolerance for checking singularity. It provides options to specify the following:
  - The type of analysis for each model effect: Type I, Type III or both
  - A value for starting iteration for checking complete and quasi-complete separation
  - The confidence interval level for coefficient estimates and estimated marginal means
  - Parameter estimate covariance matrix: Model-based estimator or robust estimator
  - The Hessian convergence criterion
  - Initial values for parameter estimates
  - Log-likelihood convergence criterion
  - Form of the log-likelihood function
  - Maximum number of iterations for parameter estimation and log-likelihood
  - Maximum number of steps in step-halving method
  - Model parameters estimation method: Fisher scoring method or Newton-Raphson method
  - Parameter convergence criterion
  - Method of fitting the scale parameter: Maximum likelihood, deviance, Pearson Chi-square or fixed at a number
  - Tolerance value used to test for singularity

- The REPEATED subcommand specifies the working correlation matrix structure used by GEE to model correlations within subjects, and controls statistical criteria in the non-likelihood-based iterative fitting

algorithm. It provides options to specify the following:

- The within-subject or time effect
- Correlation matrix structure: Independent working correlation matrix, AR(1) working correlation matrix, exchangeable working correlation matrix, fixed working correlation matrix, m-dependent working correlation matrix and unstructured working correlation matrix
- Whether to adjust the working correlation matrix estimator by the number of non-redundant parameters
- Whether to use the robust or the model-based estimator or the parameter estimate covariance matrix for generalized estimating equations
- The Hessian convergence criterion for the generalized estimating equations
- Maximum iterations
- Relative or absolute parameter convergence criterion
- The number of iterations between updates of the working correlation matrix
- To display estimated marginal means of the dependent variable for all level combinations of a set of factors

• The EMMEANS subcommand displays estimated marginal means of the dependent variable for all level combinations of a set of factors. It offers the option to specify the following:
- The cells for which estimated marginal means are displayed

- The covariate values to use when computing the estimated marginal means
- Whether to compute estimated marginal means based on the original scale of the dependent variable or on the link function transformation
- The factor or set of crossed factors, the levels or level combinations which are compared using the contrast type specified by using the CONTRAST keyword
- The type of contrast to use for the levels of the factor, or level combinations of the crossed factors, by using the COMPARE keyword. The following contrast types are available: Pairwise, deviation, difference, Helmert, polynomial, repeated and simple.
- The method of adjusting the significance level used in tests of the contrasts: Least significant difference, Bonferroni, Sequential Bonferroni, Sidak and Sequential

• The MISSING subcommand specifies how missing values are handled
• The PRINT subcommand offers options to display the following: Correlation matrix for parameter estimates, covariance matrix for parameter estimates, case processing summary, descriptive statistics, goodness of fit, general estimable function, iteration history, Lagrange multiplier test, set of contrast coefficient (L) matrices, model information, parameter estimates and corresponding statistics, model summary statistics and working correlation matrix

• The SAVE subcommand offers options to save the following to the working data file: Predicted value of the linear predictor, estimated standard error of the predicted value of the linear predictor, predicted value of the mean of the response, confidence interval for the mean of the response, leverage value, raw residual, Pearson residual, deviance residual, standardized Pearson residual, standardized deviance residual, likelihood residual and Cook's distance
• The OUTFILE subcommand offers options to save the following to an external file: The parameter correlation matrix and other statistics to an IBM SPSS Statistics dataset, the parameter covariance matrix and other statistics to an IBM SPSS Statistics dataset, and the parameter estimates and the parameter covariance matrix to an XML file
• GENLIN: check HCONVERGE after convergence even if it's not specified

MIXED
Expands the general linear model used in the GLM procedure so that data can exhibit correlation and non-constant variability
• Fit the following types of models:
- Fixed effects ANOVA model, randomized complete blocks design, split-plot design, purely random effects model, random coefficient model, multilevel analysis, unconditional linear growth model, linear growth model with person-level covariate, repeated measures analysis and repeated measures analysis with time-dependent covariate

- Opt to apply frequency weights or regression weights
- 99Use one of six covariance structures offered: First-Order autoregressive, compound asymmetry, Huynh-Feldt, identity, unstructured and variance components
- Select from 11 non-spatial covariance types: First-order ante-dependence, heterogeneous, first-order autoregressive, ARMA (1,1), heterogeneous compound symmetry, compound symmetry with correlation parameterization, diagonal, first-order factor analytic, Toeplitz, heterogeneous Toeplitz and unstructured correlations
- Choose CRITERIA to control the iterative algorithm used in estimation and to specify numerical tolerance for checking singularity: Confidence interval level, log-likelihood function convergence criterion, maximum number of iterations, parameter estimates convergence criterion (absolute and relative), maximum step-halving allowed, apply scoring algorithm, and value used as tolerance in checking singularity
- Specifiy the fixed effects in the mixed model: No intercept, Type I sum of squares and Type III sum of squares
- Specify the random effects: Identify the subjects and covariance structure (first-order autoregressive, compound symmetry, Huynh-Feldt, identity and unstructured variance components)
- Depending on the covariance type specified, random effects specified in one RANDOM subcommand may be correlated
- Use one of two estimation methods: Maximum likelihood and restricted maximum likelihood
- Select from a variety of print options: Asymptotic correlation matrix of the fixed effects parameter estimates, asymptotic covariance matrix of the fixed-effects parameter estimates, case

processing summary, descriptive statistics, estimated covariance matrix of random effects, iteration history, estimable functions, estimated covariance matrix of residual, solution for fixed-effects and random-effects parameters and tests for covariance parameters

- Use the REPEATED subcommand to specify the residual covariance matrix in the mixed effects model: Identify the subjects and covariance structure (first-order autoregressive, compound symmetry, Huynh-Feldt, identity, unstructured and variance components)
- Save fixed predicted values, predicted values, and residuals
- Use the TEST subcommand to customize hypotheses tests by directly specifying null hypotheses as linear combinations of parameters
  - Supply divisor for coefficients of random effects

- Save standard error of prediction
- Means subcommand for fixed effects, which displays the dependent variable's estimated marginal means in the cells and its standard errors for the specified factors

GLM
Describe the relationship between a dependent variable and a set of independent variables
- Select univariate and multivariate lack-of-fit tests
- Regression model
- Fixed effect ANOVA, ANCOVA, MANOVA and MANCOVA
- Random or mixed ANOVA and ANCOVA
- Repeated measures: Univariate or multivariate
- Doubly multivariate design
- Four types of sums of squares
- Full-parameterization approach to estimate parameters in the model

- General linear hypothesis testing for parameters in the model
- Write a covariance or correlation matrix of the parameter estimates in the model in a matrix data file
- Plots: Spread vs. level, residual and profile
- GLM Repeated Measures dialogs offer "contrast none" option that has no effect on output
- Post hoc tests for observed cell means: Student-Newman-Keuls, Tukey's honestly significant difference, Tukey's b, Duncan's multiple comparison procedure based on the Studentized range test, Scheffé's multiple comparison t test, Dunnett's one-tailed t test (compares if the mean at any level is smaller than that of the reference category), Dunnett's two-tailed t test (compares if the mean at any level is larger than that of the reference category), Bonferroni t test, least significant difference t test, Sidak t test, Hochberg's GT2, Gabriel's pairwise comparisons test based on the Studentized maximum modulus test, Ryan-Einot-Gabriel-Welsch's multiple stepdown procedure based on an F test, Ryan-Einot-Gabriel-Welsch's multiple stepdown procedure based on the Studentized range test, Tamhane's T2, Dunnett's T3, Games and Howell's pairwise comparisons test based on the Studentized range test, Dunnett's C and Waller-Duncan t test
- User-specified error term in post hoc analysis
- Estimated population marginal means for predicted cell means
- Save variables to the active file: Unstandardized predicted values, weighted unstandardized predicted values, unstandardized residuals, weighted unstandardized residuals, deleted residuals, standardized residuals, Studentized residuals, standard errors of predicted value,

Cook's distance and uncentered leverage values
- Fractional numbers in LMATRIX, MMATRIX and KMATRIX subcommands
- Pairwise comparisons of expected marginal means
- Linear hypothesis testing of an effect vs. a linear combination of effects
- Option to save design matrices
- Contrasts: Deviations, simple, difference, Helmert, polynomial, repeated and special
- Print: Descriptive statistics, tests of homogeneity of variance, parameter estimates, partial Eta^2, general estimable function table, lack-of-fit tests, observed power for each test, and a set of contrast coefficient (L) matrices

VARCOMP
Variance component estimation
- Estimation methods: ANOVA MINQUE, maximum likelihood (ML) and restricted maximum likelihood (REML)
- Type I and Type III sums of squares for the ANOVA method
- Choices of zero-weight or uniform-weight methods
- Choices of ML and REML calculation methods: Fisher's scoring method or Newton-Raphson method
- Save variance components estimates and covariance matrices
- Criteria specification: Iterations, convergence and Epsilon value used as tolerance in checking singularity
- Print: Expected mean squares, iteration history and sums of squares

SURVIVAL
Analysis of life tables
- Life tables for individual groups
- Interval variable lengths
- Plots: Cumulative survival distribution on log or linear scale, hazard function and density function

- Comparisons of subgroups
- Plots of the one-minus survival function
- Status variables to indicate if the terminal event occurred for the observation
- Print life tables
- Calculate comparisons of the subgroups: Exact, approximate, conditional, pairwise and compare
- Option to write survival table data records and label records files

LOGLINEAR
General models of multiway contingency tables (syntax only)
- ML estimation
- Models: Saturated, hierarchical or nonhierarchical single degree of freedom partitions and logit
- Observed and expected frequencies
- Raw and standardized residuals
- Parameter estimates
- Cell weight and structural zero specification
- Plots of adjusted residual vs. observed/expected counts
- Normal and de-trended probability plots of adjusted residuals
- Likelihood ratio and Pearson Chi-squares
- Contrasts: Deviation, difference, Helmert, simple, repeated, polynomial and special

HILOGLINEAR
Hierarchical loglinear models for multiway contingency tables
- Simultaneous entry and backward elimination methods
- Print: Frequencies and residuals
- Parameter estimates and partial associations for saturated models
- Criteria specification: Convergence, maximum iterations, probability of Chi-square for model and maximum steps
- Specified cell weights and maximum order of terms

- Plots of standardized residuals vs. observed and expected counts
- Normal probability plots of standardized residuals
- Pivot table output

GENLOG
Fit loglinear and logit models to count data by means of a generalized linear model approach
- Model fit, using ML estimation under Poisson loglinear model and multinomial loglinear models
- Exponential of the Beta
- GLM approach handles "messy data"
- Cell structure specification
- Model designs are specified through GLM model syntax
- Accommodate structural zeros
- Print Chi-square goodness-of-fit statistics
- Generalized log-odds ratio facility tests whether the specific generalized log-odds ratios are equal to zero, and can print confidence intervals
- Cell statistics include expected cell counts, residual, standardized, adjusted and deviance residual
- Include generalized residuals facility
- Diagnostic plots include high-resolution scatterplots and normal probability plots of residual statistics
- Print parameter estimates, along with correlations and covariances of the estimates
- Save residuals, standardized residuals, adjusted residuals, deviance residuals and predicted values
- Criteria specification: Confidence interval, iterations, convergence, Delta and Epsilon values used as tolerance in checking singularity

KAPLAN-MEIER

Estimates the length of time to an event using Kaplan-Meier estimation methods

- Define factors and strata
- Plots: Cumulative hazard functions, cumulative and log survival
- Display censored cases
- Save variables to a file: Cumulative number of events, hazard, standard error and survival function
- Statistical display: Cumulative events and survival, mean and median survival times with standard errors, number at risk, requested percentiles and standard error
- Tests for equality of survival distributions: Breslow, Tarone and logrank
- Specify a trend component for factor levels having a metric
- Include plots of the one-minus survival function
- Status variables to indicate if the terminal event occurred for the observation
- Specify strata (subgroups) within categories of factors
- Compare the survival distributions for different levels of the factor: Compare all factor levels in a single test, compare each pair of factors, pool the test statistic across all strata, and compare the factor levels for each stratum

COX REGRESSION

Proportional hazards with time-dependent covariates

- Contrasts: Deviations, simple, difference, Helmert, polynomial, repeated, special and indicator
- Define strata to estimate separate baseline functions

- Methods: Backward and forward stepwise and direct entry
- Plots: Cumulative survival, hazard, and log-minus-log plots for each stratum
- Removal of variables: Change in likelihood ratio, conditional and Wald
- Save variables to files: Baseline survival and hazard functions and their standard errors, cumulative hazard function, dfbeta, log-minus-log of survival function, residuals and survival function
- Include plots of the one-minus survival function
- Status variables to indicate if the terminal event occurred for the observation
- Specify ordinal or nominal predictors
- Print: Full regression output including overall model statistics for variables in the equation and variables not in the equation, summary information, correlation/covariance matrix of the parameter estimates for the variables in the model, baseline table, and confidence intervals for exponential of Beta
- Criteria: Change in parameter estimates for terminating iteration; maximum number of iterations; percentage of change in log-likelihood ratio for terminating iteration; probability of score statistic for variable entry; and probability of Wald, likelihood ratio (LR), or conditional LR statistic to remove a variable
- Specify the pattern of covariate values to be used for requested plots and coefficient tables
- Write to external IBM SPSS Statistics data files: Coefficients in the final model and survival table

System requirements

- Requirements vary according to platform. For details, see **www.spss. com/statistics.**

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, predictive analytics, financial performance and strategy management, and analytic applications provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises – able to direct and automate decisions to meet business goals and achieve measurable competitive advantage.

Business Analytics software