IBM SPSS Regression 28

Note

Before using this information and the product it supports, read the information in “Notices” on page 31.

Product Information

This edition applies to version 28, release 0, modification 0 of IBM® SPSS® Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

© Copyright International Business Machines Corporation .
US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Chapter 1. Regression 1

Choosing a procedure for Binary Logistic Regression 1
Logistic Regression 2
Logistic Regression Set Rule 3
Logistic Regression Variable Selection Methods 3
Logistic Regression Define Categorical Variables 3
Logistic Regression Save New Variables 4
Logistic Regression Options 5
LOGISTIC REGRESSION Command Additional Features 5
Multinomial Logistic Regression 5
Multinomial Logistic Regression 6
Multinomial Logistic Regression Reference Category 7
Multinomial Logistic Regression Statistics 7
Multinomial Logistic Regression Criteria 8
Multinomial Logistic Regression Options 8
Multinomial Logistic Regression Save 9
NOMREG Command Additional Features 9
Probit Regression 9
Probit Analysis Define Range 10
Probit Analysis Options 10
PROBIT Command Additional Features 11
Quantile Regression 11
Quantile Regression: Criteria 12
Quantile Regression: Model 13
Quantile Regression: Display 14
Quantile Regression: Save 15
Quantile Regression: Export 16
Nonlinear Regression 17
Conditional Logic (Nonlinear Regression) 17
Nonlinear Regression Parameters 18
Nonlinear Regression Common Models 18
Nonlinear Regression Loss Function 19
Nonlinear Regression Parameter Constraints 19
Nonlinear Regression Save New Variables 20
Nonlinear Regression Options 20
Interpreting Nonlinear Regression Results 20
NLR Command Additional Features 20
Weight Estimation 21
Weight Estimation Options 22
WLS Command Additional Features 22
Two-Stage Least-Squares Regression 22
Two-Stage Least-Squares Regression Options 23
2SLS Command Additional Features 23
Categorical Variable Coding Schemes 23
Deviation 24
Simple 24
Helmert 25
Difference 25
Polynomial 25
Repeated 26
Special 26

Indicator 27
Kernel Ridge Regression 27
Kernel Parameters 28
Kernel Ridge Regression: Options 29

Notices 31
Trademarks 32

Index 33

Chapter 1. Regression

The following regression features are included in SPSS Statistics Standard Edition or the Regression option.

Choosing a procedure for Binary Logistic Regression

Binary logistic regression models can be fitted using the Logistic Regression procedure and the Multinomial Logistic Regression procedure. Each procedure has options not available in the other. An important theoretical distinction is that the Logistic Regression procedure produces all predictions, residuals, influence statistics, and goodness-of-fit tests using data at the individual case level, regardless of how the data are entered and whether or not the number of covariate patterns is smaller than the total number of cases, while the Multinomial Logistic Regression procedure internally aggregates cases to form subpopulations with identical covariate patterns for the predictors, producing predictions, residuals, and goodness-of-fit tests based on these subpopulations. If all predictors are categorical or any continuous predictors take on only a limited number of values—so that there are several cases at each distinct covariate pattern—the subpopulation approach can produce valid goodness-of-fit tests and informative residuals, while the individual case level approach cannot.

Logistic Regression

Provides the following unique features:

- Hosmer-Lemeshow test of goodness of fit for the model
- Stepwise analyses
- Contrasts to define model parameterization
- Alternative cut points for classification
- Classification plots
- Model fitted on one set of cases to a held-out set of cases
- Saves predictions, residuals, and influence statistics

Multinomial Logistic Regression

Provides the following unique features:

- Pearson and deviance chi-square tests for goodness of fit of the model
- Specification of subpopulations for grouping of data for goodness-of-fit tests
- Listing of counts, predicted counts, and residuals by subpopulations
- Correction of variance estimates for over-dispersion
- Covariance matrix of the parameter estimates
- Tests of linear combinations of parameters
- Explicit specification of nested models
- Fit 1-1 matched conditional logistic regression models using differenced variables

Notes:

- Both of these procedures fit a model for binary data that is a generalized linear model with a binomial distribution and logit link function. If a different link function is more appropriate for your data, then you should use the Generalized Linear Models procedure.
- If you have repeated measurements of binary data, or records that are otherwise correlated, then you should consider the Generalized Linear Mixed Models or Generalized Estimating Equations procedures.

Logistic Regression

Logistic regression is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous. Logistic regression coefficients can be used to estimate odds ratios for each of the independent variables in the model. Logistic regression is applicable to a broader range of research situations than discriminant analysis.

Example. What lifestyle characteristics are risk factors for coronary heart disease (CHD)? Given a sample of patients measured on smoking status, diet, exercise, alcohol use, and CHD status, you could build a model using the four lifestyle variables to predict the presence or absence of CHD in a sample of patients. The model can then be used to derive estimates of the odds ratios for each factor to tell you, for example, how much more likely smokers are to develop CHD than nonsmokers.

Statistics. For each analysis: total cases, selected cases, valid cases. For each categorical variable: parameter coding. For each step: variable(s) entered or removed, iteration history, –2 log-likelihood, goodness of fit, Hosmer-Lemeshow goodness-of-fit statistic, model chi-square, improvement chi-square, classification table, correlations between variables, observed groups and predicted probabilities chart, residual chi-square. For each variable in the equation: coefficient (*B*), standard error of *B*, Wald statistic, estimated odds ratio (exp(*B*)), confidence interval for exp(*B*), log-likelihood if term removed from model. For each variable not in the equation: score statistic. For each case: observed group, predicted probability, predicted group, residual, standardized residual.

Methods. You can estimate models using block entry of variables or any of the following stepwise methods: forward conditional, forward LR, forward Wald, backward conditional, backward LR, or backward Wald.

Logistic Regression data considerations

Data. The dependent variable should be dichotomous. Independent variables can be interval level or categorical; if categorical, they should be dummy or indicator coded (there is an option in the procedure to recode categorical variables automatically).

Assumptions. Logistic regression does not rely on distributional assumptions in the same sense that discriminant analysis does. However, your solution may be more stable if your predictors have a multivariate normal distribution. Additionally, as with other forms of regression, multicollinearity among the predictors can lead to biased estimates and inflated standard errors. The procedure is most effective when group membership is a truly categorical variable; if group membership is based on values of a continuous variable (for example, "high IQ" versus "low IQ"), you should consider using linear regression to take advantage of the richer information offered by the continuous variable itself.

Related procedures. Use the Scatterplot procedure to screen your data for multicollinearity. If assumptions of multivariate normality and equal variance-covariance matrices are met, you may be able to get a quicker solution using the Discriminant Analysis procedure. If all of your predictor variables are categorical, you can also use the Loglinear procedure. If your dependent variable is continuous, use the Linear Regression procedure. You can use the ROC Curve procedure to plot probabilities saved with the Logistic Regression procedure.

Obtaining a Logistic Regression Analysis

1. From the menus choose:

   Analyze > **Regression** > **Binary Logistic...**

2. Select one dichotomous dependent variable. This variable may be numeric or string.
3. Select one or more covariates. To include interaction terms, select all of the variables involved in the interaction and then select **>a*b>**.

To enter variables in groups (**blocks**), select the covariates for a block, and click **Next** to specify a new block. Repeat until all blocks have been specified.

Optionally, you can select cases for analysis. Choose a selection variable, and enter the rule criteria.

Logistic Regression Set Rule

Cases defined by the selection rule are included in model estimation. For example, if you selected a variable and **equals** and specified a value of 5, then only the cases for which the selected variable has a value equal to 5 are included in estimating the model.

Statistics and classification results are generated for both selected and unselected cases. This provides a mechanism for classifying new cases based on previously existing data, or for partitioning your data into training and testing subsets, to perform validation on the model generated.

Logistic Regression Variable Selection Methods

Method selection allows you to specify how independent variables are entered into the analysis. Using different methods, you can construct a variety of regression models from the same set of variables.

- *Enter*. A procedure for variable selection in which all variables in a block are entered in a single step.
- *Forward Selection (Conditional)*. Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on conditional parameter estimates.
- *Forward Selection (Likelihood Ratio)*. Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates.
- *Forward Selection (Wald)*. Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of the Wald statistic.
- *Backward Elimination (Conditional)*. Backward stepwise selection. Removal testing is based on the probability of the likelihood-ratio statistic based on conditional parameter estimates.
- *Backward Elimination (Likelihood Ratio)*. Backward stepwise selection. Removal testing is based on the probability of the likelihood-ratio statistic based on the maximum partial likelihood estimates.
- *Backward Elimination (Wald)*. Backward stepwise selection. Removal testing is based on the probability of the Wald statistic.

The significance values in your output are based on fitting a single model. Therefore, the significance values are generally invalid when a stepwise method is used.

All independent variables selected are added to a single regression model. However, you can specify different entry methods for different subsets of variables. For example, you can enter one block of variables into the regression model using stepwise selection and a second block using forward selection. To add a second block of variables to the regression model, click **Next**.

Logistic Regression Define Categorical Variables

You can specify details of how the Logistic Regression procedure will handle categorical variables:

Covariates. Contains a list of 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 last category) is compared to the next category.
- **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.

Logistic Regression Save New Variables

You can save results of the logistic regression as new variables in the active dataset:

Predicted Values. Saves values predicted by the model. Available options are Probabilities and Group membership.

- *Probabilities*. For each case, saves the predicted probability of occurrence of the event. A table in the output displays name and contents of any new variables. The "event" is the category of the dependent variable with the higher value; for example, if the dependent variable takes values 0 and 1, the predicted probability of category 1 is saved.
- *Predicted Group Membership*. The group with the largest posterior probability, based on discriminant scores. The group the model predicts the case belongs to.

Influence. Saves values from statistics that measure the influence of cases on predicted values. Available options are Cook's, Leverage values, and DfBeta(s).

- *Cook's*. The logistic regression analog of Cook's influence statistic. 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.
- *Leverage Value*. The relative influence of each observation on the model's fit.
- *DfBetas*. The difference in beta value is the change in the regression coefficient that results from the exclusion of a particular case. A value is computed for each term in the model, including the constant.

Residuals. Saves residuals. Available options are Unstandardized, Logit, Studentized, Standardized, and Deviance.

- *Unstandardized Residuals*. The difference between an observed value and the value predicted by the model.
- *Logit Residual*. The residual for the case if it is predicted in the logit scale. The logit residual is the residual divided by the predicted probability times 1 minus the predicted probability.
- *Studentized Residual*. The change in the model deviance if a case is excluded.
- *Standardized Residuals*. 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.
- *Deviance*. Residuals based on the model deviance.

Export model information to XML file. Parameter estimates and (optionally) their covariances are exported to the specified file in XML (PMML) format. You can use this model file to apply the model information to other data files for scoring purposes.

Logistic Regression Options

You can specify options for your logistic regression analysis:

Statistics and Plots. Allows you to request statistics and plots. Available options are Classification plots, Hosmer-Lemeshow goodness-of-fit, Casewise listing of residuals, Correlations of estimates, Iteration history, and CI for exp(*B*). Select one of the alternatives in the Display group to display statistics and plots either At each step or, only for the final model, At last step.

- *Hosmer-Lemeshow goodness-of-fit statistic*. This goodness-of-fit statistic is more robust than the traditional goodness-of-fit statistic used in logistic regression, particularly for models with continuous covariates and studies with small sample sizes. It is based on grouping cases into deciles of risk and comparing the observed probability with the expected probability within each decile.

Probability for Stepwise. Allows you to control the criteria by which variables are entered into and removed from the equation. You can specify criteria for Entry or Removal of variables.

- *Probability for Stepwise*. A variable is entered into the model if the probability of its score statistic is less than the Entry value and is removed if the probability is greater than the Removal value. To override the default settings, enter positive values for Entry and Removal. Entry must be less than Removal.

Classification cutoff. Allows you to determine the cut point for classifying cases. Cases with predicted values that exceed the classification cutoff are classified as positive, while those with predicted values smaller than the cutoff are classified as negative. To change the default, enter a value between 0.01 and 0.99.

Maximum Iterations. Allows you to change the maximum number of times that the model iterates before terminating.

Include constant in model. Allows you to indicate whether the model should include a constant term. If disabled, the constant term will equal 0.

LOGISTIC REGRESSION Command Additional Features

The command syntax language also allows you to:

- Identify casewise output by the values or variable labels of a variable.
- Control the spacing of iteration reports. Rather than printing parameter estimates after every iteration, you can request parameter estimates after every *n*th iteration.
- Change the criteria for terminating iteration and checking for redundancy.
- Specify a variable list for casewise listings.
- Conserve memory by holding the data for each split file group in an external scratch file during processing.

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

Multinomial Logistic Regression

Multinomial Logistic regression is useful for situations in which you want to be able to classify subjects based on values of a set of predictor variables. This type of regression is similar to logistic regression, but it is more general because the dependent variable is not restricted to two categories.

Example. In order to market films more effectively, movie studios want to predict what type of film a moviegoer is likely to see. By performing a Multinomial Logistic Regression, the studio can determine the strength of influence a person's age, gender, and dating status has upon the type of film they prefer. The studio can then slant the advertising campaign of a particular movie toward a group of people likely to go see it.

Statistics. Iteration history, parameter coefficients, asymptotic covariance and correlation matrices, likelihood-ratio tests for model and partial effects, –2 log-likelihood. Pearson and deviance chi-square goodness of fit. Cox and Snell, Nagelkerke, and McFadden R^2. Classification: observed versus predicted frequencies by response category. Crosstabulation: observed and predicted frequencies (with residuals) and proportions by covariate pattern and response category.

Methods. A multinomial logit model is fit for the full factorial model or a user-specified model. Parameter estimation is performed through an iterative maximum-likelihood algorithm.

Multinomial Logistic Regression data considerations

Data. The dependent variable should be categorical. Independent variables can be factors or covariates. In general, factors should be categorical variables and covariates should be continuous variables.

Assumptions. It is assumed that the odds ratio of any two categories are independent of all other response categories. For example, if a new product is introduced to a market, this assumption states that the market shares of all other products are affected proportionally equally. Also, given a covariate pattern, the responses are assumed to be independent multinomial variables.

Obtaining a Multinomial Logistic Regression

1. From the menus choose:

   Analyze > **Regression** > **Multinomial Logistic Regression...**
2. Select one dependent variable.
3. Factors are optional and can be either numeric or categorical.
4. Covariates are optional but must be numeric if specified.

Multinomial Logistic Regression

By default, the Multinomial Logistic Regression procedure produces a model with the factor and covariate main effects, but you can specify a custom model or request stepwise model selection with this dialog box.

Specify Model. A main-effects model contains the covariate and factor main effects but no interaction effects. A full factorial model contains all main effects and all factor-by-factor interactions. It does not contain covariate interactions. You can create a custom model to specify subsets of factor interactions or covariate interactions, or request stepwise selection of model terms.

Factors & Covariates. The factors and covariates are listed.

Forced Entry Terms. Terms added to the forced entry list are always included in the model.

Stepwise Terms. Terms added to the stepwise list are included in the model according to one of the following user-selected Stepwise Methods:

- **Forward entry.** This method begins with no stepwise terms in the model. At each step, the most significant term is added to the model until none of the stepwise terms left out of the model would have a statistically significant contribution if added to the model.
- **Backward elimination.** This method begins by entering all terms specified on the stepwise list into the model. At each step, the least significant stepwise term is removed from the model until all of the remaining stepwise terms have a statistically significant contribution to the model.
- **Forward stepwise.** This method begins with the model that would be selected by the forward entry method. From there, the algorithm alternates between backward elimination on the stepwise terms in the model and forward entry on the terms left out of the model. This continues until no terms meet the entry or removal criteria.
- **Backward stepwise.** This method begins with the model that would be selected by the backward elimination method. From there, the algorithm alternates between forward entry on the terms left out of the model and backward elimination on the stepwise terms in the model. This continues until no terms meet the entry or removal criteria.

Include intercept in model. Allows you to include or exclude an intercept term for the model.

Build Terms

For the selected factors and covariates:

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.

Multinomial Logistic Regression Reference Category

By default, the Multinomial Logistic Regression procedure makes the last category the reference category. This dialog box gives you control of the reference category and the way in which categories are ordered.

Reference Category. Specify the first, last, or a custom category.

Category Order. In ascending order, the lowest value defines the first category and the highest value defines the last. In descending order, the highest value defines the first category and the lowest value defines the last.

Multinomial Logistic Regression Statistics

You can specify the following statistics for your Multinomial Logistic Regression:

Case processing summary. This table contains information about the specified categorical variables.

Model. Statistics for the overall model.

- **Pseudo R-square.** Prints the Cox and Snell, Nagelkerke, and McFadden R^2 statistics.
- **Step summary.** This table summarizes the effects entered or removed at each step in a stepwise method. It is not produced unless a stepwise model is specified in the Model dialog box.
- **Model fitting information.** This table compares the fitted and intercept-only or null models.
- **Information criteria.** This table prints Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC).
- **Cell probabilities.** Prints a table of the observed and expected frequencies (with residual) and proportions by covariate pattern and response category.
- **Classification table.** Prints a table of the observed versus predicted responses.
- **Goodness of fit chi-square statistics.** Prints Pearson and likelihood-ratio chi-square statistics. Statistics are computed for the covariate patterns determined by all factors and covariates or by a user-defined subset of the factors and covariates.
- **Monotinicity measures.** Displays a table with information on the number of concordant pairs, discordant pairs, and tied pairs. The Somers' D, Goodman and Kruskal's Gamma, Kendall's tau-a, and Concordance Index C are also displayed in this table.

Parameters. Statistics related to the model parameters.

- **Estimates.** Prints estimates of the model parameters, with a user-specified level of confidence.
- **Likelihood ratio test.** Prints likelihood-ratio tests for the model partial effects. The test for the overall model is printed automatically.
- **Asymptotic correlations.** Prints matrix of parameter estimate correlations.
- **Asymptotic covariances.** Prints matrix of parameter estimate covariances.

Define Subpopulations. Allows you to select a subset of the factors and covariates in order to define the covariate patterns used by cell probabilities and the goodness-of-fit tests.

Multinomial Logistic Regression Criteria

You can specify the following criteria for your Multinomial Logistic Regression:

Iterations. Allows you to specify the maximum number of times you want to cycle through the algorithm, the maximum number of steps in the step-halving, the convergence tolerances for changes in the log-likelihood and parameters, how often the progress of the iterative algorithm is printed, and at what iteration the procedure should begin checking for complete or quasi-complete separation of the data.

- **Log-likelihood convergence.** Convergence is assumed if the absolute change in the log-likelihood function is less than the specified value. The criterion is not used if the value is 0. Specify a non-negative value.
- **Parameter convergence.** Convergence is assumed if the absolute change in the parameter estimates is less than this value. The criterion is not used if the value is 0.

Delta. Allows you to specify a non-negative value less than 1. This value is added to each empty cell of the crosstabulation of response category by covariate pattern. This helps to stabilize the algorithm and prevent bias in the estimates.

Singularity tolerance. Allows you to specify the tolerance used in checking for singularities.

Multinomial Logistic Regression Options

You can specify the following options for your Multinomial Logistic Regression:

Dispersion Scale. Allows you to specify the dispersion scaling value that will be used to correct the estimate of the parameter covariance matrix. **Deviance** estimates the scaling value using the deviance function (likelihood-ratio chi-square) statistic. **Pearson** estimates the scaling value using the Pearson chi-square statistic. You can also specify your own scaling value. It must be a positive numeric value.

Stepwise Options. These options give you control of the statistical criteria when stepwise methods are used to build a model. They are ignored unless a stepwise model is specified in the Model dialog box.

- **Entry Probability.** This is the probability of the likelihood-ratio statistic for variable entry. The larger the specified probability, the easier it is for a variable to enter the model. This criterion is ignored unless the forward entry, forward stepwise, or backward stepwise method is selected.
- **Entry test.** This is the method for entering terms in stepwise methods. Choose between the likelihood-ratio test and score test. This criterion is ignored unless the forward entry, forward stepwise, or backward stepwise method is selected.
- **Removal Probability.** This is the probability of the likelihood-ratio statistic for variable removal. The larger the specified probability, the easier it is for a variable to remain in the model. This criterion is ignored unless the backward elimination, forward stepwise, or backward stepwise method is selected.
- **Removal Test.** This is the method for removing terms in stepwise methods. Choose between the likelihood-ratio test and Wald test. This criterion is ignored unless the backward elimination, forward stepwise, or backward stepwise method is selected.
- **Minimum Stepped Effects in Model.** When using the backward elimination or backward stepwise methods, this specifies the minimum number of terms to include in the model. The intercept is not counted as a model term.
- **Maximum Stepped Effects in Model.** When using the forward entry or forward stepwise methods, this specifies the maximum number of terms to include in the model. The intercept is not counted as a model term.
- **Hierarchically constrain entry and removal of terms.** This option allows you to choose whether to place restrictions on the inclusion of model terms. Hierarchy requires that for any term to be included, all lower order terms that are a part of the term to be included must be in the model first. For example, if the hierarchy requirement is in effect, the factors *Marital status* and *Gender* must both be in the model

before the *Marital status*Gender* interaction can be added. The three radio button options determine the role of covariates in determining hierarchy.

Multinomial Logistic Regression Save

The Save dialog box allows you to save variables to the working file and export model information to an external file.

Saved variables. The following variables can be saved:

- **Estimated response probabilities.** These are the estimated probabilities of classifying a factor/covariate pattern into the response categories. There are as many estimated probabilities as there are categories of the response variable; up to 25 will be saved.
- **Predicted category.** This is the response category with the largest expected probability for a factor/covariate pattern.
- **Predicted category probabilities.** This is the maximum of the estimated response probabilities.
- **Actual category probability.** This is the estimated probability of classifying a factor/covariate pattern into the observed category.

Export model information to XML file. Parameter estimates and (optionally) their covariances are exported to the specified file in XML (PMML) format. You can use this model file to apply the model information to other data files for scoring purposes.

NOMREG Command Additional Features

The command syntax language also allows you to:

- Specify the reference category of the dependent variable.
- Include cases with user-missing values.
- Customize hypothesis tests by specifying null hypotheses as linear combinations of parameters.

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

Probit Regression

This procedure measures the relationship between the strength of a stimulus and the proportion of cases exhibiting a certain response to the stimulus. It is useful for situations where you have a dichotomous output that is thought to be influenced or caused by levels of some independent variable(s) and is particularly well suited to experimental data. This procedure will allow you to estimate the strength of a stimulus required to induce a certain proportion of responses, such as the median effective dose.

Example. How effective is a new pesticide at killing ants, and what is an appropriate concentration to use? You might perform an experiment in which you expose samples of ants to different concentrations of the pesticide and then record the number of ants killed and the number of ants exposed. Applying probit regression to these data, you can determine the strength of the relationship between concentration and killing, and you can determine what the appropriate concentration of pesticide would be if you wanted to be sure to kill, say, 95% of exposed ants.

Statistics. Regression coefficients and standard errors, intercept and standard error, Pearson goodness-of-fit chi-square, observed and expected frequencies, and confidence intervals for effective levels of independent variable(s). Plots: transformed response plots.

This procedure uses the algorithms proposed and implemented in NPSOL® by Gill, Murray, Saunders & Wright to estimate the model parameters.

Probit Regression data considerations

Data. For each value of the independent variable (or each combination of values for multiple independent variables), your response variable should be a count of the number of cases with those values that show the response of interest, and the total observed variable should be a count of the total number of

cases with those values for the independent variable. The factor variable should be categorical, coded as integers.

Assumptions. Observations should be independent. If you have a large number of values for the independent variables relative to the number of observations, as you might in an observational study, the chi-square and goodness-of-fit statistics may not be valid.

Related procedures. Probit analysis is closely related to logistic regression; in fact, if you choose the logit transformation, this procedure will essentially compute a logistic regression. In general, probit analysis is appropriate for designed experiments, whereas logistic regression is more appropriate for observational studies. The differences in output reflect these different emphases. The probit analysis procedure reports estimates of effective values for various rates of response (including median effective dose), while the logistic regression procedure reports estimates of odds ratios for independent variables.

Obtaining a Probit Regression analysis

1. From the menus choose:

   Analyze > **Regression** > **Probit...**

2. Select a response frequency variable. This variable indicates the number of cases exhibiting a response to the test stimulus. The values of this variable cannot be negative.
3. Select a total observed variable. This variable indicates the number of cases to which the stimulus was applied. The values of this variable cannot be negative and cannot be less than the values of the response frequency variable for each case.

   Optionally, you can select a Factor variable. If you do, use **Define Range** to define the range for the groups.
4. Select one or more covariate(s). This variable contains the level of the stimulus applied to each observation. If you want to transform the covariate, select a transformation from the **Transform** drop-down list. If no transformation is applied and there is a control group, then the control group is included in the analysis.
5. Select either the **Probit** or **Logit** model.

   Probit Model
   Applies the probit transformation (the inverse of the cumulative standard normal distribution function) to the response proportions.

   Logit Model
   Applies the logit (log odds) transformation to the response proportions.

Probit Analysis Define Range

This allows you to specify the levels of the factor variable that will be analyzed. The factor levels must be coded as consecutive integers, and all levels in the range that you specify will be analyzed.

Probit Analysis Options

You can specify options for your probit analysis:

Statistics. Allows you to request the following optional statistics: Frequencies, Relative median potency, Parallelism test, and Fiducial confidence intervals.

- *Relative Median Potency*. Displays the ratio of median potencies for each pair of factor levels. Also shows 95% confidence limits for each relative median potency. Relative median potencies are not available if you do not have a factor variable or if you have more than one covariate.
- *Parallelism Test*. A test of the hypothesis that all factor levels have a common slope.
- *Fiducial Confidence Intervals*. Confidence intervals for the dosage of agent required to produce a certain probability of response.

Fiducial confidence intervals and Relative median potency are unavailable if you have selected more than one covariate. Relative median potency and Parallelism test are available only if you have selected a factor variable.

Natural Response Rate. Allows you to indicate a natural response rate even in the absence of the stimulus. Available alternatives are None, Calculate from data, or Value.

- *Calculate from Data*. Estimate the natural response rate from the sample data. Your data should contain a case representing the control level, for which the value of the covariates is 0. Probit estimates the natural response rate using the proportion of responses for the control level as an initial value.
- *Value*. Sets the natural response rate in the model (select this item when you know the natural response rate in advance). Enter the natural response proportion (the proportion must be less than 1). For example, if the response occurs 10% of the time when the stimulus is 0, enter 0.10.

Criteria. Allows you to control parameters of the iterative parameter-estimation algorithm. You can override the defaults for Maximum iterations, Step limit, and Optimality tolerance.

PROBIT Command Additional Features

The command syntax language also allows you to:

- Request an analysis on both the probit and logit models.
- Control the treatment of missing values.
- Transform the covariates by bases other than base 10 or natural log.

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

Quantile Regression

Regression is a statistical method broadly used in quantitative modeling. Multiple linear regression is a basic and standard approach in which researchers use the values of several variables to explain or predict the mean values of a scale outcome. However, in many circumstances, we are more interested in the median, or an arbitrary quantile of the scale outcome.

Quantile regression models the relationship between a set of predictor (independent) variables and specific percentiles (or "quantiles") of a target (dependent) variable, most often the median. It has two main advantages over Ordinary Least Squares regression:

- Quantile regression makes no assumptions about the distribution of the target variable.
- Quantile regression tends to resist the influence of outlying observations

Quantile regression is widely used for researching in industries such as ecology, healthcare, and financial economics.

Example

What is the relationship between total household income and the proportion of income that is spent on food? Engel's law is an observation in economics stating that as income rises, the proportion of income spent on food falls, even if absolute expenditure on food rises. Applying quantile regression to these data, you can determine which food expense can cover 90% of families (for 100 families with a given income) when not interested in the mean food expense.

Statistics

Quantile Regression, Simplex approach, Frisch-Newton interior-point non-linear optimization algorithm, Barrodale and Roberts, Bofinger, Hall Sheather, bandwidth, significance level, matrix manipulations, convergence criterion, regression weights, intercept term, predicted target, prediction residuals, tabulation, prediction plots, parameter estimates, covariance matrix, correlation matrix, observed values, confidence interval.

This procedure uses the algorithms proposed by Koenker, R. W. and Bassett, G. W. (1978). Regression quantiles, Econometrica, 46, 33–50.

Quantile Regression data considerations

Data
A single numeric dependent variable is required. The target variable needs to be a continuous variable. The predictors can be continuous variables or dummy variables for categorical predictors. Either the intercept term or at least one predictor is required to run an analysis.

Assumptions
Quantile regression does not make assumptions on the distribution of the target variable and resists the influence of outlying observations.

Related procedures
Quantile analysis is related to Ordinary Least Squares regression.

Obtaining a Quantile Regression analysis

1. From the menus choose:

   Analyze > **Regression** > **Quantile...**

   The dialog allows you to specify the target, factor, covariate, and weight variables to use for quantile regression analysis. The dialog also provides the option of conserving memory for complex analysis or large datasets.
2. Select a numeric target variable. Only one target variable is required to run an analysis. Only numeric variables are allowed.
3. Optionally, select one or more factor variables. Scale variables are not allowed.
4. Optionally, select one or more covariate variables. String variables are not allowed.

   Note: When both the **Factor(s)** and **Covariate(s)** lists are empty, and **Include intercept in model** is selected on the Model dialog, the following message displays:

   ```
   No effects have been specified. Therefore, an intercept only model will be fit.
   Do you want to fit an intercept-only model?
   ```
5. Optionally, select a regression weight variable. String variables are not allowed.
6. Optionally, select **Conserve memory for complex analysis or large datasets**. This setting controls whether or not the data is held in an external file during processing. Enabling the setting can help conserve memory resources when running complex analyses, or analyses with large data sets.

Quantile Regression: Criteria

The Criteria dialog provides options for

Quantile
Provides options for specifying the quantile(s).

Specify single quantiles
When selected, at least one value is required to run the analysis. Multiple values are allowed and each value must belong to `[0, 1]`. You can specify multiple values with each value separated by a blank space (or blank spaces). Use the **Add**, **Change**, and **Remove** buttons to work with the values in the quantile value list.

All values must be unique (duplicate values are not allowed). The default value is `0.5`.

Specify grid quantiles
When selected, a grid of quantiles can be specified from a **Start** value (`value1`) to an **End** value (`value2`) with the increment of **By** (`value3`). If specified, only one valid set of [`value1 TO value2 BY value3`] is allowed. It must satisfy that $0 \leq value1 \leq value2 \leq 1$. In cases where $value1 = value2$, it is equivalent to specifying a single `value1`, regardless of `value3`.

Estimation Method
Provides options for specifying the model estimation method.

Automatically chosen by the program
Allows the procedure to automatically select the appropriate estimation method. This is the default setting.

Simplex algorithm
Calls the simplex algorithm that was developed by Barrodale and Roberts.

Frisch-Newton interior-point non-linear optimization
Calls for the Frisch-Newton interior-point non-linear optimization algorithm.

Post-estimation
Provides options for the post-estimation of the variance-covariance of the parameter estimates and the confidence intervals for the predicted target values.

Assume cases are IID
When selected, this setting assumes that error terms are independently and identically distributed. When the setting is not selected, the computation time may significantly increase for large models. The setting is selected by default.

Bandwidth type
Determines which bandwidth method is used to estimate the variance-covariance matrix of the parameter estimates (**Bofinger** or **Hall-Sheather**). **Bofinger** is the default setting.

Numerical Method
Provides the following options:

Singularity tolerance
Specifies the tolerance value for the matrix manipulations in the interior-point method. The specified value must be a single, double value in $(0, 10^{-3})$, with 10^{-12} as the default setting.

Convergence
Specifies the convergence criterion for the numerical method. The specified value must be a single, double value in $(0, 10^{-3})$, with 10^{-6} as the default setting.

Maximum iterations
Specifies the maximum number of iterations. The specified value must be a single, positive integer. The default value is 2000.

Missing Values
Provides options for determining how missing values are handled.

Exclude both user-missing and system missing values
When selected, both user-missing and system missing values are excluded.

User-missing values are treated as valid
When selected, user-missing values are treated as valid.

Confidence interval (%)
Specifies the significance level. When specified, the value must be a single double value in between 0 and 100. The default value is 95.

Quantile Regression: Model

The Model dialog provides options for specifying the effects and the weights that are used in the model. If omitted, or specified by itself, the model will contain the intercept term and all main effects with the covariates in the covariates list and the factors in factors list.

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. When **Build terms** is selected, the following effect and interaction options are available for non-nested terms.

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.

When **Build nested terms** is selected, you can build nested terms. 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.

Nesting notes:

- To include an effect for an interaction between two factors, use the keyword BY or the asterisk (*) to join the factors that are involved in the interaction.
- Factors inside an interaction effect must be distinct.
- Use parenthesis pairs to include an effect for nesting one term within another.
- When more than one pair of parentheses are present, each pair of parentheses must be enclosed or nested within another pair of parentheses.
- Multiple nesting is allowed.
- Interactions between nested effects are not supported.

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.

Include intercept in model
When selected, the intercept term is included in the model. When not selected, at least one predictor is required to run the analysis. The setting is enabled by default.

Quantile Regression: Display

The Display dialog provides output and plot control settings.

Print
The following output options are available.

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.

Plot and Tabulate

The following plotting options are available:

Plot the parameter estimates for

You can select to plot parameter estimates for a specific number of top effects, or for all effects in the model. **Top xx effects** setting controls the number of categories, or combinations of categories in a mixed effect, that are plotted within the interaction of one covariate and one or two factors. The value must be a single positive integer (50 is the default setting).

Notes:

- Prediction plots are created for all effects when the integer value you specify is larger than the number of categories or combinations.
- The setting is valid only when multiple values are specified for the **Quantile values** setting on the Criteria dialog. No plots are created when a single quantile value is specified.

Display the predicted by observed plot

Controls the creation of the predicted versus observed values plot. When enabled, a single plot that contains the points (with different colored dots representing different quantiles) is created. The setting is disabled by default.

Predict the effects in the model

When enabled, the following options are available:

Plot or tabulate the top x effects

Specify the number of top effects whose prediction plot or prediction table will be created. 3 is the default value.

Note: The prediction plots or prediction tables are created for all effects when the specified value is larger than the number of valid effects in the model.

Plot or tabulate user-specified effects

Valid effect guidelines are:

- Effect with one covariate (including a high power of the covariate itself): Create a single plot containing the lines predicted by different quantiles.
- Effect with one factor: Tabulate the predictions for the categories of the factor by different quantiles.
- Effect with the interaction of two factors: For each quantile, tabulate the predictions for the categories of two factors.
- Effect with the interaction of one covariate and one or two factors: For each quantile, create a plot containing the lines for each category or combination of the categories within the interaction effect.
- The maximum number of the combinations to be plotted is controlled by the value specified for **Plot maximum xx categories of combinations of categories in a mixed effect**.

Effects that are moved from **Model Effects** to the **Prediction Lines** list are used for plotting. Plots are not created in cases where the specified effects are constant (removed from model building).

Effects that are moved from **Model Effects** to the **Prediction Tables** list are used for tabulation. Tables are not created in cases where the specified effects are constant (removed from model building).

Plot maximum xx categories or combinations of categories in a mixed effect

Controls the maximum number of the category combinations to plot. The default value is 10.

Quantile Regression: Save

The Save dialog provides options for scoring the model.

Predicted value of response

When selected, predicted target value are scored.

Residual
When selected, prediction residuals are scored.

Lower bound of prediction interval
When selected, the lower bounds of the prediction intervals are scored.

Upper bound of prediction interval
When selected, the upper bounds of the prediction intervals are scored.

Note: A variable name can be specified for each save option. If a root name is specified, it must be a valid variable name. A root name, followed by an underscore "_" character and a meaningful quantile suffix, is used when multiple values are specified for the **Quantile values** setting on the Criteria dialog.

Quantile Regression: Export

The Export dialog provides options for specifying which statistics are exported, how statistics are exported (external data file or data sets), and controlling how data is handled during processing (process normally, or held in an external scratch file while processing).

Covariance matrix of parameter estimates
When selected, options for writing the covariance matrix of the parameter estimates to an external data file, or a previously declared data, set are enabled.

Correlation matrix of parameter estimates
When selected, options for writing the correlation matrix of the parameter estimates to an external data file, or a previously declared data set, are enabled.

The covariance/correlation matrix will be saved in a single dataset or external file in the presence of multiple regression quantiles
When multiple quantiles are present, this option toggles the saving of covariance/correlation matrices to single or multiple data sets or external data files. When not enabled, matrices are saved in a single, external data file or a data set. When enabled, matrices are saved in multiple external data files or data sets. The setting takes effect only when multiple values are specified for the **Quantile values** setting on the Criteria dialog.

Note: This option is available only when **Covariance matrix of parameter estimates** or **Correlation matrix of parameter estimates** is selected.

Export model information to XML file
When selected, provides options for exporting the model information to a specific XML file name and location.

Export as XML
When **Export model information to XML file** is selected, you can select to export either parameter estimates and covariance matrices or parameter estimates only. **Parameter estimates and covariance matrix** is the default setting.

File name conventions

- When a single value is specified for the **Quantile values** setting on the Criteria dialog, `savefile` and `dataset` are used to name the external data file or data set.
- When multiple values are specified for the **Quantile values** setting on the Criteria dialog, each quantile is saved to an external data file or data set.
- An underscore character "_", followed by a meaningful quantile suffix, is automatically appended to the data file or data set name. For example, when 0.25, 0.50, and 0.75 are specified as **Quantile values**, the suffix `_25`, `_50`, and `_75` are appended to the data file names (before the `.sav` extension).
- Additional digits can be specified for each quantile suffix (if necessary).
- The **Quantile values** leading zero and decimal point are not used in the suffix.
- When scientific notation is specified for **Quantile values**, it is converted to a decimal value when displayed in the suffix.

Nonlinear Regression

Nonlinear regression is a method of finding a nonlinear model of the relationship between the dependent variable and a set of independent variables. Unlike traditional linear regression, which is restricted to estimating linear models, nonlinear regression can estimate models with arbitrary relationships between independent and dependent variables. This is accomplished using iterative estimation algorithms. Note that this procedure is not necessary for simple polynomial models of the form Y = A + BX**2. By defining W = X**2, we get a simple linear model, Y = A + BW, which can be estimated using traditional methods such as the Linear Regression procedure.

Example. Can population be predicted based on time? A scatterplot shows that there seems to be a strong relationship between population and time, but the relationship is nonlinear, so it requires the special estimation methods of the Nonlinear Regression procedure. By setting up an appropriate equation, such as a logistic population growth model, we can get a good estimate of the model, allowing us to make predictions about population for times that were not actually measured.

Statistics. For each iteration: parameter estimates and residual sum of squares. For each model: sum of squares for regression, residual, uncorrected total and corrected total, parameter estimates, asymptotic standard errors, and asymptotic correlation matrix of parameter estimates.

Note: Constrained nonlinear regression uses the algorithms proposed and implemented in NPSOL® by Gill, Murray, Saunders, and Wright to estimate the model parameters.

Nonlinear Regression data considerations

Data. The dependent and independent variables should be quantitative. Categorical variables, such as religion, major, or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables.

Assumptions. Results are valid only if you have specified a function that accurately describes the relationship between dependent and independent variables. Additionally, the choice of good starting values is very important. Even if you've specified the correct functional form of the model, if you use poor starting values, your model may fail to converge or you may get a locally optimal solution rather than one that is globally optimal.

Related procedures. Many models that appear nonlinear at first can be transformed to a linear model, which can be analyzed using the Linear Regression procedure. If you are uncertain what the proper model should be, the Curve Estimation procedure can help to identify useful functional relations in your data.

Obtaining a Nonlinear Regression Analysis

1. From the menus choose:

   Analyze > **Regression** > **Nonlinear...**

2. Select one numeric dependent variable from the list of variables in your active dataset.
3. To build a model expression, enter the expression in the **Model Expression** field or paste components (variables, parameters, functions) into the field.
4. Identify parameters in your model by clicking **Parameters**.

A segmented model (one that takes different forms in different parts of its domain) must be specified by using conditional logic within the single model statement.

Conditional Logic (Nonlinear Regression)

You can specify a segmented model using conditional logic. To use conditional logic within a model expression or a loss function, you form the sum of a series of terms, one for each condition. Each term consists of a logical expression (in parentheses) multiplied by the expression that should result when that logical expression is true.

For example, consider a segmented model that equals 0 for X<=0, X for 0<X<1, and 1 for X>=1. The expression for this is:

(X<=0)*0 + (X>0 & X<1)*X + (X>=1)*1.

The logical expressions in parentheses all evaluate to 1 (true) or 0 (false). Therefore:

If X<=0, the above reduces to 1*0 + 0*X + 0*1 = 0.

If 0<X<1, it reduces to 0*0 + 1*X + 0*1 = X.

If X>=1, it reduces to 0*0 + 0*X + 1*1 = 1.

More complicated examples can be easily built by substituting different logical expressions and outcome expressions. Remember that double inequalities, such as 0<X<1, must be written as compound expressions, such as (X>0 & X<1).

String variables can be used within logical expressions:

(city='New York')*costliv + (city='Des Moines')*0.59*costliv

This yields one expression (the value of the variable *costliv*) for New Yorkers and another (59% of that value) for Des Moines residents. String constants must be enclosed in quotation marks or apostrophes, as shown here.

Nonlinear Regression Parameters

Parameters are the parts of your model that the Nonlinear Regression procedure estimates. Parameters can be additive constants, multiplicative coefficients, exponents, or values used in evaluating functions. All parameters that you have defined will appear (with their initial values) on the Parameters list in the main dialog box.

Name. You must specify a name for each parameter. This name must be a valid variable name and must be the name used in the model expression in the main dialog box.

Starting Value. Allows you to specify a starting value for the parameter, preferably as close as possible to the expected final solution. Poor starting values can result in failure to converge or in convergence on a solution that is local (rather than global) or is physically impossible.

Use starting values from previous analysis. If you have already run a nonlinear regression from this dialog box, you can select this option to obtain the initial values of parameters from their values in the previous run. This permits you to continue searching when the algorithm is converging slowly. (The initial starting values will still appear on the Parameters list in the main dialog box.)

Note: This selection persists in this dialog box for the rest of your session. If you change the model, be sure to deselect it.

Nonlinear Regression Common Models

The table below provides example model syntax for many published nonlinear regression models. A model selected at random is not likely to fit your data well. Appropriate starting values for the parameters are necessary, and some models require constraints in order to converge.

Table 1. Example model syntax

Name	Model expression
Asymptotic Regression	b1 + b2 * exp(b3 * x)
Asymptotic Regression	b1 – (b2 * (b3 ** x))
Density	(b1 + b2 * x) ** (–1 / b3)
Gauss	b1 * (1 – b3 * exp(–b2 * x ** 2))
Gompertz	b1 * exp(–b2 * exp(–b3 * x))
Johnson-Schumacher	b1 * exp(–b2 / (x + b3))
Log-Modified	(b1 + b3 * x) ** b2

Table 1. Example model syntax (continued)

Name	Model expression
Log-Logistic	b1 – ln(1 + b2 * exp(–b3 * x))
Metcherlich Law of Diminishing Returns	b1 + b2 * exp(–b3 * x)
Michaelis Menten	b1 * x / (x + b2)
Morgan-Mercer-Florin	(b1 * b2 + b3 * x ** b4) / (b2 + x ** b4)
Peal-Reed	b1 / (1+ b2 * exp(–(b3 * x + b4 * x **2 + b5 * x ** 3)))
Ratio of Cubics	(b1 + b2 * x + b3 * x ** 2 + b4 * x ** 3) / (b5 * x ** 3)
Ratio of Quadratics	(b1 + b2 * x + b3 * x ** 2) / (b4 * x ** 2)
Richards	b1 / ((1 + b3 * exp(–b2 * x)) ** (1 / b4))
Verhulst	b1 / (1 + b3 * exp(–b2 * x))
Von Bertalanffy	(b1 ** (1 – b4) – b2 * exp(–b3 * x)) ** (1 / (1 – b4))
Weibull	b1 – b2 * exp(–b3 * x ** b4)
Yield Density	(b1 + b2 * x + b3 * x ** 2) ** (–1)

Nonlinear Regression Loss Function

The **loss function** in nonlinear regression is the function that is minimized by the algorithm. Select either **Sum of squared residuals** to minimize the sum of the squared residuals or **User-defined loss function** to minimize a different function.

If you select **User-defined loss function**, you must define the loss function whose sum (across all cases) should be minimized by the choice of parameter values.

- Most loss functions involve the special variable *RESID_*, which represents the residual. (The default Sum of squared residuals loss function could be entered explicitly as `RESID_**2`.) If you need to use the predicted value in your loss function, it is equal to the dependent variable minus the residual.
- It is possible to specify a conditional loss function using conditional logic.

You can either type an expression in the User-defined loss function field or paste components of the expression into the field. String constants must be enclosed in quotation marks or apostrophes, and numeric constants must be typed in American format, with the dot as a decimal delimiter.

Nonlinear Regression Parameter Constraints

A **constraint** is a restriction on the allowable values for a parameter during the iterative search for a solution. Linear expressions are evaluated before a step is taken, so you can use linear constraints to prevent steps that might result in overflows. Nonlinear expressions are evaluated after a step is taken.

Each equation or inequality requires the following elements:

- An expression involving at least one parameter in the model. Type the expression or use the keypad, which allows you to paste numbers, operators, or parentheses into the expression. You can either type in the required parameter(s) along with the rest of the expression or paste from the Parameters list at the left. You cannot use ordinary variables in a constraint.
- One of the three logical operators <=, =, or >=.
- A numeric constant, to which the expression is compared using the logical operator. Type the constant. Numeric constants must be typed in American format, with the dot as a decimal delimiter.

Nonlinear Regression Save New Variables

You can save a number of new variables to your active data file. Available options are Residuals, Predicted values, Derivatives, and Loss function values. These variables can be used in subsequent analyses to test the fit of the model or to identify problem cases.

- *Residuals*. Saves residuals with the variable name resid.
- *Predicted Values*. Saves predicted values with the variable name pred_.
- *Derivatives*. One derivative is saved for each model parameter. Derivative names are created by prefixing 'd.' to the first six characters of parameter names.
- *Loss Function Values*. This option is available if you specify your own loss function. The variable name loss_ is assigned to the values of the loss function.

Nonlinear Regression Options

Options allow you to control various aspects of your nonlinear regression analysis:

Bootstrap Estimates. A method of estimating the standard error of a statistic using repeated samples from the original data set. This is done by sampling (with replacement) to get many samples of the same size as the original data set. The nonlinear equation is estimated for each of these samples. The standard error of each parameter estimate is then calculated as the standard deviation of the bootstrapped estimates. Parameter values from the original data are used as starting values for each bootstrap sample. This requires the sequential quadratic programming algorithm.

Estimation Method. Allows you to select an estimation method, if possible. (Certain choices in this or other dialog boxes require the sequential quadratic programming algorithm.) Available alternatives include Sequential quadratic programming and Levenberg-Marquardt.

- *Sequential Quadratic Programming*. This method is available for constrained and unconstrained models. Sequential quadratic programming is used automatically if you specify a constrained model, a user-defined loss function, or bootstrapping. You can enter new values for Maximum iterations and Step limit, and you can change the selection in the drop-down lists for Optimality tolerance, Function precision, and Infinite step size.
- *Levenberg-Marquardt*. This is the default algorithm for unconstrained models. The Levenberg-Marquardt method is not available if you specify a constrained model, a user-defined loss function, or bootstrapping. You can enter new values for Maximum iterations, and you can change the selection in the drop-down lists for Sum-of-squares convergence and Parameter convergence.

Interpreting Nonlinear Regression Results

Nonlinear regression problems often present computational difficulties:

- The choice of initial values for the parameters influences convergence. Try to choose initial values that are reasonable and, if possible, close to the expected final solution.
- Sometimes one algorithm performs better than the other on a particular problem. In the Options dialog, select the other algorithm if it is available. (If you specify a loss function or certain types of constraints, you cannot use the Levenberg-Marquardt algorithm.)
- When iteration stops only because the maximum number of iterations has occurred, the "final" model is probably not a good solution. Select **Use starting values from previous analysis** in the Parameters dialog to continue the iteration or, better yet, choose different initial values.
- Models that require exponentiation of or by large data values can cause overflows or underflows (numbers too large or too small for the computer to represent). Sometimes you can avoid these by suitable choice of initial values or by imposing constraints on the parameters.

NLR Command Additional Features

The command syntax language also allows you to:

- Name a file from which to read initial values for parameter estimates.

- Specify more than one model statement and loss function. This makes it easier to specify a segmented model.
- Supply your own derivatives rather than use those calculated by the program.
- Specify the number of bootstrap samples to generate.
- Specify additional iteration criteria, including setting a critical value for derivative checking and defining a convergence criterion for the correlation between the residuals and the derivatives.

Additional criteria for the CNLR (constrained nonlinear regression) command allow you to:

- Specify the maximum number of minor iterations allowed within each major iteration.
- Set a critical value for derivative checking.
- Set a step limit.
- Specify a crash tolerance to determine if initial values are within their specified bounds.

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

Weight Estimation

Standard linear regression models assume that variance is constant within the population under study. When this is not the case (for example, when cases that are high on some attribute show more variability than cases that are low on that attribute) linear regression using ordinary least squares (OLS) no longer provides optimal model estimates. If the differences in variability can be predicted from another variable, the Weight Estimation procedure can compute the coefficients of a linear regression model using weighted least squares (WLS), such that the more precise observations (that is, those with less variability) are given greater weight in determining the regression coefficients. The Weight Estimation procedure tests a range of weight transformations and indicates which will give the best fit to the data.

Example. What are the effects of inflation and unemployment on changes in stock prices? Because stocks with higher share values often show more variability than those with low share values, ordinary least squares will not produce optimal estimates. Weight estimation allows you to account for the effect of share price on the variability of price changes in calculating the linear model.

Statistics. Log-likelihood values for each power of the weight source variable tested, multiple *R*, *R*-squared, adjusted *R*-squared, ANOVA table for WLS model, unstandardized and standardized parameter estimates, and log-likelihood for the WLS model.

Weight Estimation data considerations

Data. The dependent and independent variables should be quantitative. Categorical variables, such as religion, major, or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables. The weight variable should be quantitative and should be related to the variability in the dependent variable.

Assumptions. For each value of the independent variable, the distribution of the dependent variable must be normal. The relationship between the dependent variable and each independent variable should be linear, and all observations should be independent. The variance of the dependent variable can vary across levels of the independent variable(s), but the differences must be predictable based on the weight variable.

Related procedures. The Explore procedure can be used to screen your data. Explore provides tests for normality and homogeneity of variance, as well as graphical displays. If your dependent variable seems to have equal variance across levels of independent variables, you can use the Linear Regression procedure. If your data appear to violate an assumption (such as normality), try transforming them. If your data are not related linearly and a transformation does not help, use an alternate model in the Curve Estimation procedure. If your dependent variable is dichotomous (for example, whether a particular sale is completed or whether an item is defective) use the Logistic Regression procedure. If your dependent variable is censored (for example, survival time after surgery) use Life Tables, Kaplan-Meier, or Cox Regression, available in Custom Tables and Advanced Statistics. If your data are not independent

(for example, if you observe the same person under several conditions) use the Repeated Measures procedure, available in Custom Tables and Advanced Statistics.

Obtaining a Weight Estimation Analysis

1. From the menus choose:

   Analyze > **Regression** > **Weight Estimation...**
2. Select one dependent variable.
3. Select one or more independent variables.
4. Select the variable that is the source of heteroscedasticity as the weight variable.

Weight Variable
The data are weighted by the reciprocal of this variable raised to a power. The regression equation is calculated for each of a specified range of power values and indicates the power that maximizes the log-likelihood function.

Power Range
This is used in conjunction with the weight variable to compute weights. Several regression equations will be fit, one for each value in the power range. The values entered in the Power range test box and the through text box must be between -6.5 and 7.5, inclusive. The power values range from the low to high value, in increments determined by the value specified. The total number of values in the power range is limited to 150.

Weight Estimation Options

You can specify options for your weight estimation analysis:

Save best weight as new variable. Adds the weight variable to the active file. This variable is called *WGT_n*, where *n* is a number chosen to give the variable a unique name.

Display ANOVA and Estimates. Allows you to control how statistics are displayed in the output. Available alternatives are For best power and For each power value.

WLS Command Additional Features

The command syntax language also allows you to:

- Provide a single value for the power.
- Specify a list of power values, or mix a range of values with a list of values for the power.

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

Two-Stage Least-Squares Regression

Standard linear regression models assume that errors in the dependent variable are uncorrelated with the independent variable(s). When this is not the case (for example, when relationships between variables are bidirectional), linear regression using ordinary least squares (OLS) no longer provides optimal model estimates. Two-stage least-squares regression uses instrumental variables that are uncorrelated with the error terms to compute estimated values of the problematic predictor(s) (the first stage), and then uses those computed values to estimate a linear regression model of the dependent variable (the second stage). Since the computed values are based on variables that are uncorrelated with the errors, the results of the two-stage model are optimal.

Example. Is the demand for a commodity related to its price and consumers' incomes? The difficulty in this model is that price and demand have a reciprocal effect on each other. That is, price can influence demand and demand can also influence price. A two-stage least-squares regression model might use consumers' incomes and lagged price to calculate a proxy for price that is uncorrelated with the measurement errors in demand. This proxy is substituted for price itself in the originally specified model, which is then estimated.

Statistics. For each model: standardized and unstandardized regression coefficients, multiple *R*, R^2, adjusted R^2, standard error of the estimate, analysis-of-variance table, predicted values, and residuals. Also, 95% confidence intervals for each regression coefficient, and correlation and covariance matrices of parameter estimates.

Two-Stage Least-Squares Regression data considerations

Data. The dependent and independent variables should be quantitative. Categorical variables, such as religion, major, or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables. *Endogenous* explanatory variables should be quantitative (not categorical).

Assumptions. For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and each independent variable should be linear.

Related procedures. If you believe that none of your predictor variables is correlated with the errors in your dependent variable, you can use the Linear Regression procedure. If your data appear to violate one of the assumptions (such as normality or constant variance), try transforming them. If your data are not related linearly and a transformation does not help, use an alternate model in the Curve Estimation procedure. If your dependent variable is dichotomous, such as whether a particular sale is completed or not, use the Logistic Regression procedure. If your data are not independent--for example, if you observe the same person under several conditions--use the Repeated Measures procedure.

Obtaining a Two-Stage Least-Squares Regression Analysis

1. From the menus choose:

   Analyze > **Regression** > **2-Stage Least Squares...**
2. Select one dependent variable.
3. Select one or more explanatory (predictor) variables.
4. Select one or more instrumental variables.

   - *Instrumental*. These are the variables used to compute the predicted values for the endogenous variables in the first stage of two-stage least squares analysis. The same variables may appear in both the Explanatory and Instrumental list boxes. The number of instrumental variables must be at least as many as the number of explanatory variables. If all explanatory and instrumental variables listed are the same, the results are the same as results from the Linear Regression procedure.

Explanatory variables not specified as instrumental are considered endogenous. Normally, all of the exogenous variables in the Explanatory list are also specified as instrumental variables.

Two-Stage Least-Squares Regression Options

You can select the following options for your analysis:

Save New Variables. Allows you to add new variables to your active file. Available options are Predicted and Residuals.

Display covariance of parameters. Allows you to print the covariance matrix of the parameter estimates.

2SLS Command Additional Features

The command syntax language also allows you to estimate multiple equations simultaneously. See the *Command Syntax Reference* for complete syntax information.

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.

Kernel Ridge Regression

Kernel Ridge Regression is an extension procedure that uses the Python **sklearn.kernel_ridge.KernelRidge** class to estimate kernel ridge regression models. Kernel ridge regression models are nonparametric regression models that are capable of modeling linear and nonlinear relationships between predictor variables and outcomes. Results can be highly sensitive to choices of model hyperparameters. Kernel Ridge Regression facilitates choice of hyperparameter values through k-fold cross-validation on specified grids of values using the **sklearn.model_selection.GridSearchCV** class.

Example

Statistics

Additive_CHI2, CHI2, Cosine, Laplacian, Linear, Polynomial, RBF, Sigmoid, Alpha, Gamma, Coef0, Degree, crossvalidation, observed versus predicted, residuals versus predicted, dual weight coefficients, kernel space weight coefficients.

Data considerations

Data

- You can specify any or all of the eight different kernel functions.
- The selected kernel function determines which hyperparameters are active.
- Hyperparameters include alpha for ridge regularization that are common to all kernels plus as many as three other hyperparameters for each specific kernel function.
- When multiple kernel subcommands are specified, or more than one value for any parameter is specified, a grid search with cross-validation to evaluate models is performed, and the best fitting model that is based on held out data is selected.
- The extension accepts split variables from the Split File procedure and weights using the Weight Cases procedure.
- When weights are included, they are used in creating fitted values in all analyses. Due to limitations in the score method in the **sklearn.model_selection.GridSearchCV** class, crossvalidation evaluations that are used for model selection are not weighted.

Assumptions

Obtaining a Kernel Ridge Regression

1. From the menus choose:

   Analyze > **Regression** > **Kernel Ridge...**

2. Select a **Dependent** variable.

3. Select one or more **Independent(s)** variables.
4. The default **Single model** setting is used when only one value for each kernel function parameter is specified. When the **Single model** setting is selected, you cannot specify additional **Kernel(s)** functions and weights are fully applied throughout the analysis, evaluation, and scoring of results. You can also use the up and down arrow controls to rearrange the kernel functions.

   Optionally, select **Model selection** from the **Mode** list.

   When **Model selection** is selected from the **Mode** list, you can add multiple kernel functions to the **Kernel(s)** list.

   a. Click the add control (**+**) to include additional kernel functions.

   b. Click the empty cell in the **Kernel** column to select a kernel function.

   c. Double-click any kernel function row cell to specify kernel function parameter values for the corresponding column (**Alpha**, **Gamma**, **Coef0**, **Degree**). For more information, see "Kernel Parameters" on page 28. The default kernel function tuning parameters are listed below.

   Additive_CHI2
   ALPHA=1 GAMMA=1

   CHI2
   ALPHA=1 GAMMA=1

   Cosine
   ALPHA=1

   Laplacian
   ALPHA=1 GAMMA=1/p

   Linear
   The default kernel function. ALPHA=1

   Polynomial
   ALPHA=1 GAMMA=1/p COEF0=1 DEGREE=3

   RBF
   ALPHA=1 GAMMA=1/p

   Sigmoid
   ALPHA=1 GAMMA=1/p COEF0=1

   Note: When more than one value for any kernel function parameter is specified, a grid search with cross-validation to evaluate models is performed, and the best fitting model that is based on held out data is selected.
5. Optionally, click **Options** to specify the number of crossvalidation folds, display options, plot settings, and items to save. For more information, see "Kernel Ridge Regression: Options" on page 29.
6. Click **OK**.

Kernel Parameters

The **Kernel Parameters** dialog provides options for specifying single kernel function parameter values and for specifying that model selection is performed using a grid search over the combinations of kernels and specified grid parameter values.

Specify single parameters
Enable the setting to specify values for the selected kernel function parameter.

- Enter a value and click **Add** to include the value in the kernel function parameter.
- Select a parameter value and click **Change** to update the value.
- Select a parameter value and click **Remove** to delete the value.

Specify grid parameters
Enable the setting to specify that model selection is performed using a grid search over the combinations of kernels and specified grid parameter values.

Kernel Ridge Regression: Options

The **Plots** dialog provides options for specifying the number of crossvalidation folds, display options, plot settings, and items to save.

Number of crossvalidation folds
The number of splits or folds in crossvalidation with grid search for model selection. Enter an integer value larger than 1. The default value is 5. The setting is available only when **Model selection** is chosen as the **Mode** on the primary **Kernel Ridge Regression** dialog.

Display
Provides options for specifying which output to display when crossvalidation is in effect.

Best
The default setting displays only basic results for the chosen best model.

Compare
Displays basic results for all evaluated models.

Compare models and folds
Displays full results for each split or fold for each evaluated model.

Plot
Provides options for specifying plots of observed or residual values versus predicted values.

Observed vs. Predicted
Displays a scatterplot of observed versus predicted values for the specified or best model.

Residuals vs. Predicted
Displays a scatterplot of residuals versus predicted values for the specified or best model.

Save
The table provides options for specifying variables to save to the active dataset.

Predicted values
Saves predicted values from the specified or best model to the active dataset. An optional variable name can be included.

Residuals
Saves residuals from the specified or best model predictions to the active dataset. An optional variable name can be included.

Dual coefficients
Saves dual or kernel space weight coefficients from the specified model to the active dataset. An optional variable name can be included. The setting is not available when **Model selection** is chosen as the **Mode** on the primary **Kernel Ridge Regression** dialog.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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, MD-NC119
Armonk, NY 10504-1785
US

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.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

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 jurisdictions 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 websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

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

Licensees of this program who wish 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 Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

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.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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-IBMproducts. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

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 actual people or business enterprises 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:

© Copyright IBM Corp. 2021. Portions of this code are derived from IBM Corp. Sample Programs.

© Copyright IBM Corp. 1989 - 2021. 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

A

asymptotic regression
  in Nonlinear Regression 18

B

backward elimination
  in Logistic Regression 3
binary logistic regression 1, 2

C

categorical covariates 3
cell probabilities tables
  in Multinomial Logistic Regression 7
cells with zero observations
  in Multinomial Logistic Regression 8
classification
  in Multinomial Logistic Regression 5
classification tables
  in Multinomial Logistic Regression 7
confidence intervals
  in Multinomial Logistic Regression 7
constant term
  in Linear Regression 5
constrained regression
  in Nonlinear Regression 19
contrasts
  in Logistic Regression 3
convergence criterion
  in Multinomial Logistic Regression 8
Cook's D
  in Logistic Regression 4
correlation matrix
  in Multinomial Logistic Regression 7
covariance matrix
  in Multinomial Logistic Regression 7
covariates
  in Logistic Regression 3
Cox and Snell R-square
  in Multinomial Logistic Regression 7
custom models
  in Multinomial Logistic Regression 6

D

delta
  as correction for cells with zero observations 8
density model
  in Nonlinear Regression 18
deviance function
  for estimating dispersion scaling value 8
DfBeta
  in Logistic Regression 4
dispersion scaling value
dispersion scaling value *(continued)*
  in Multinomial Logistic Regression 8

F

fiducial confidence intervals
  in Probit Analysis 10
forward selection
  in Logistic Regression 3
full factorial models
  in Multinomial Logistic Regression 6

G

Gauss model
  in Nonlinear Regression 18
Gompertz model
  in Nonlinear Regression 18
goodness of fit
  in Multinomial Logistic Regression 7

H

Hosmer-Lemeshow goodness-of-fit statistic
  in Logistic Regression 5

I

intercept
  include or exclude 6
iteration history
  in Multinomial Logistic Regression 8
iterations
  in Logistic Regression 5
  in Multinomial Logistic Regression 8
  in Probit Analysis 10

J

Johnson-Schumacher model
  in Nonlinear Regression 18

K

Kernel Ridge
  alpha 27
  coef0 27
  degree 27
  gamma 27
  model selection 27
  single model 27
Kernel Ridge Regression
  crossvalidation folds 29
  display 29
  grid parameters 28
  parameters 28

Kernel Ridge Regression *(continued)*
plots 29
save 29

L

leverage values
in Logistic Regression 4
likelihood ratio
for estimating dispersion scaling value 8
goodness of fit 7
Linear Regression
Two-Stage Least-Squares Regression 22
weight estimation 21
log-likelihood
in Multinomial Logistic Regression 7
in Weight Estimation 21
log-modified model
in Nonlinear Regression 18
logistic regression 2
Logistic Regression
binary 1
categorical covariates 3
classification cutoff 5
coefficients 2
command additional features 5
constant term 5
contrasts 3
define selection rule 3
display options 5
example 2
Hosmer-Lemeshow goodness-of-fit statistic 5
influence measures 4
iterations 5
predicted values 4
probability for stepwise 5
residuals 4
saving new variables 4
set rule 3
statistics 2
statistics and plots 5
string covariates 3
variable selection methods 3
logistic regression analysis 2

M

main-effects models
in Multinomial Logistic Regression 6
McFadden R-square
in Multinomial Logistic Regression 7
Metcherlich law of diminishing returns
in Nonlinear Regression 18
Michaelis Menten model
in Nonlinear Regression 18
Morgan-Mercer-Florin model
in Nonlinear Regression 18
Multinomial Logistic Regression
command additional features 9
criteria 8
exporting model information 9
models 6
Multinomial Logistic Regression *(continued)*
reference category 7
save 9
statistics 7

N

Nagelkerke R-square
in Multinomial Logistic Regression 7
nonlinear models
in Nonlinear Regression 18
Nonlinear Regression
bootstrap estimates 20
command additional features 20
common nonlinear models 18
conditional logic 17
derivatives 20
estimation methods 20
example 17
interpretation of results 20
Levenberg-Marquardt algorithm 20
loss function 19
parameter constraints 19
parameters 18
predicted values 20
residuals 20
save new variables 20
segmented model 17
sequential quadratic programming 20
starting values 18
statistics 17

P

parallelism test
in Probit Analysis 10
parameter constraints
in Nonlinear Regression 19
parameter estimates
in Multinomial Logistic Regression 7
Peal-Reed model
in Nonlinear Regression 18
Pearson chi-square
for estimating dispersion scaling value 8
goodness of fit 7
Probit Analysis
command additional features 11
criteria 10
define range 10
fiducial confidence intervals 10
iterations 10
natural response rate 10
parallelism test 10
relative median potency 10
statistics 10
Probit Regression
example 9
statistics 9

Q

Quantile Regression
criteria 12

Quantile Regression *(continued)*
display 14
example 11
export 16
model 13
save 15
statistics 11

R

ratio of cubics model
in Nonlinear Regression 18
ratio of quadratics model
in Nonlinear Regression 18
reference category
in Multinomial Logistic Regression 7
relative median potency
in Probit Analysis 10
Richards model
in Nonlinear Regression 18

S

separation
in Multinomial Logistic Regression 8
singularity
in Multinomial Logistic Regression 8
SPSS logistic regression 2
step-halving
in Multinomial Logistic Regression 8
stepwise selection
in Logistic Regression 3
in Multinomial Logistic Regression 6
string covariates
in Logistic Regression 3
survival analysis
in Kernel Ridge Regression 27

T

Two-Stage Least-Squares Regression
command additional features 23
covariance of parameters 23
example 22
instrumental variables 22
saving new variables 23
statistics 22

V

Verhulst model
in Nonlinear Regression 18
Von Bertalanffy model
in Nonlinear Regression 18

W

Weibull model
in Nonlinear Regression 18
Weight Estimation
command additional features 22
display ANOVA and estimates 22
Weight Estimation *(continued)*
example 21
iteration history 22
log-likelihood 21
save best weights as new variable 22
statistics 21

Y

yield density model
in Nonlinear Regression 18