IBM SPSS Missing Values 30

Note

Before using this information and the product it supports, read the information in "Notices" on page 19.

Contents

Chapter 1. Missing values

The following missing values features are included in SPSS Statistics Professional Edition or the Missing Values option.

Introduction to Missing Values

Cases with missing values pose an important challenge, because typical modeling procedures simply discard these cases from the analysis. When there are few missing values (very roughly, less than 5% of the total number of cases) and those values can be considered to be missing at random; that is, whether a value is missing does not depend upon other values, then the typical method of listwise deletion is relatively "safe". The Missing Values option can help you to determine whether listwise deletion is sufficient, and provides methods for handling missing values when it is not.

Missing Value Analysis versus Multiple Imputation procedures

The Missing Values option provides two sets of procedures for handling missing values:

- The Multiple Imputation procedures provide analysis of patterns of missing data, geared toward eventual multiple imputation of missing values. That is, multiple versions of the dataset are produced, each containing its own set of imputed values. When statistical analyses are performed, the parameter estimates for all of the imputed datasets are pooled, providing estimates that are generally more accurate than they would be with only one imputation.

- Missing Value Analysis provides a slightly different set of descriptive tools for analyzing missing data (most particularly Little's MCAR test), and includes a variety of single imputation methods. Note that multiple imputation is generally considered to be superior to single imputation.

Missing Values Tasks

You can get started with analysis of missing values by following these basic steps:

1. **Examine missingness.** Use Missing Value Analysis and Analyze Patterns to explore patterns of missing values in your data and determine whether multiple imputation is necessary.

2. **Impute missing values.** Use Impute Missing Data Values to multiply impute missing values.

3. **Analyze "complete" data.** Use any procedure that supports multiple imputation data. See "Analyzing Multiple Imputation Data" on page 13 for information on analyzing multiple imputation datasets and a list of procedures which support these data.

Missing Value Analysis

The Missing Value Analysis procedure performs three primary functions:

- Describes the pattern of missing data. Where are the missing values located? How extensive are they? Do pairs of variables tend to have values missing in multiple cases? Are data values extreme? Are values missing randomly?

- Estimates means, standard deviations, covariances, and correlations for different missing value methods: listwise, pairwise, regression, or EM (expectation-maximization). The pairwise method also displays counts of pairwise complete cases.

- Fills in (imputes) missing values with estimated values using regression or EM methods; however, multiple imputation is generally considered to provide more accurate results.

Missing value analysis helps address several concerns caused by incomplete data. If cases with missing values are systematically different from cases without missing values, the results can be misleading. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.

Example. In evaluating a treatment for leukemia, several variables are measured. However, not all measurements are available for every patient. The patterns of missing data are displayed, tabulated, and found to be random. An EM analysis is used to estimate the means, correlations, and covariances. It is also used to determine that the data are missing completely at random. Missing values are then replaced by imputed values and saved into a new data file for further analysis.

Statistics. Univariate statistics, including number of nonmissing values, mean, standard deviation, number of missing values, and number of extreme values. Estimated means, covariance matrix, and correlation matrix, using listwise, pairwise, EM, or regression methods. Little's MCAR test with EM results. Summary of means by various methods. For groups defined by missing versus nonmissing values: *t* tests. For all variables: missing value patterns displayed cases-by-variables.

Data Considerations

Data. Data can be categorical or quantitative (scale or continuous). However, you can estimate statistics and impute missing data only for the quantitative variables. For each variable, missing values that are not coded as system-missing must be defined as user-missing. For example, if a questionnaire item has the response *Don't know* coded as 5 and you want to treat it as missing, the item should have 5 coded as a user-missing value.

Frequency weights. Frequency (replication) weights are honored by this procedure. Cases with negative or zero replication weight value are ignored. Noninteger weights are truncated.

Assumptions. Listwise, pairwise, and regression estimation depend on the assumption that the pattern of missing values does not depend on the data values. (This condition is known as **missing completely at random**, or MCAR.) Therefore, all methods (including the EM method) for estimation give consistent and unbiased estimates of the correlations and covariances when the data are MCAR. Violation of the MCAR assumption can lead to biased estimates produced by the listwise, pairwise, and regression methods. If the data are not MCAR, you need to use EM estimation.

EM estimation depends on the assumption that the pattern of missing data is related to the observed data only. (This condition is called **missing at random**, or MAR.) This assumption allows estimates to be adjusted using available information. For example, in a study of education and income, the subjects with low education may have more missing income values. In this case, the data are MAR, not MCAR. In other words, for MAR, the probability that income is recorded depends on the subject's level of education. The probability may vary by education but not by income *within that level of education*. If the probability that income is recorded also varies by the value of income within each level of education (for example, people with high incomes don't report them), then the data are neither MCAR nor MAR. This is not an uncommon situation, and, if it applies, none of the methods is appropriate.

Related procedures. Many procedures allow you to use listwise or pairwise estimation. Linear Regression and Factor Analysis allow replacement of missing values by the mean values. In the Forecasting add-on module, several methods are available to replace missing values in time series.

To Obtain Missing Value Analysis

1. From the menus choose:

   Analyze > **Missing Value Analysis…**

2. Select at least one quantitative (scale) variable for estimating statistics and optionally imputing missing values.

Optionally, you can:

- Select categorical variables (numeric or string) and enter a limit on the number of categories (**Maximum Categories**).
- Click **Patterns** to tabulate patterns of missing data. See the topic "Displaying Patterns of Missing Values" on page 3 for more information.
- Click **Descriptives** to display descriptive statistics of missing values. See the topic "Displaying Descriptive Statistics for Missing Values" on page 4 for more information.

- Select a method for estimating statistics (means, covariances, and correlations) and possibly imputing missing values. See the topic "Estimating Statistics and Imputing Missing Values" on page 4 for more information.
- If you select EM or Regression, click **Variables** to specify a subset to be used for the estimation. See the topic "Predicted and Predictor Variables" on page 6 for more information.
- Select a case label variable. This variable is used to label cases in patterns tables that display individual cases.

Displaying Patterns of Missing Values

You can choose to display various tables showing the patterns and extent of missing data. These tables can help you identify:

- Where missing values are located
- Whether pairs of variables tend to have missing values in individual cases
- Whether data values are extreme

Display

Three types of tables are available for displaying patterns of missing data.

Tabulated cases. The missing value patterns in the analysis variables are tabulated, with frequencies shown for each pattern. Use **Sort variables by missing value pattern** to specify whether counts and variables are sorted by similarity of patterns. Use **Omit patterns with less than n % of cases** to eliminate patterns that occur infrequently.

Cases with missing values. Each case with a missing or extreme value is tabulated for each analysis variable. Use **Sort variables by missing value pattern** to specify whether counts and variables are sorted by similarity of patterns.

All cases. Each case is tabulated, and missing and extreme values are indicated for each variable. Cases are listed in the order they appear in the data file, unless a variable is specified in **Sort by**.

In the tables that display individual cases, the following symbols are used:

+. Extremely high value

-. Extremely low value

S. System-missing value

A. First type of user-missing value

B. Second type of user-missing value

C. Third type of user-missing value

Variables

You can display additional information for the variables that are included in the analysis. The variables that you add to **Additional Information for** are displayed individually in the missing patterns table. For quantitative (scale) variables, the mean is displayed; for categorical variables, the number of cases having the pattern in each category is displayed.

- **Sort by.** Cases are listed according to the ascending or descending order of the values of the specified variable. Available only for **All cases**.

To Display Missing Value Patterns

1. In the main Missing Value Analysis dialog box, select the variable(s) for which you want to display missing value patterns.
2. Click **Patterns**.
3. Select the pattern table(s) that you want to display.

Displaying Descriptive Statistics for Missing Values

Univariate Statistics

Univariate statistics can help you identify the general extent of missing data. For each variable, the following are displayed:

- Number of nonmissing values
- Number and percentage of missing values

For quantitative (scale) variables, the following are also displayed:

- Mean
- Standard deviation
- Number of extremely high and low values

Indicator Variable Statistics

For each variable, an indicator variable is created. This categorical variable indicates whether the variable is present or missing for an individual case. The indicator variables are used to create the mismatch, *t* test, and frequency tables.

Percent mismatch. For each pair of variables, displays the percentage of cases in which one variable has a missing value and the other variable has a nonmissing value. Each diagonal element in the table contains the percentage of missing values for a single variable.

t tests with groups formed by indicator variables. The means of two groups are compared for each quantitative variable, using Student's *t* statistic. The groups specify whether a variable is present or missing. The *t* statistic, degrees of freedom, counts of missing and nonmissing values, and means of the two groups are displayed. You can also display any two-tailed probabilities associated with the *t* statistic. If your analysis results in more than one test, do not use these probabilities for significance testing. The probabilities are appropriate only when a single test is calculated.

Crosstabulations of categorical and indicator variables. A table is displayed for each categorical variable. For each category, the table shows the frequency and percentage of nonmissing values for the other variables. The percentages of each type of missing value are also displayed.

Omit variables missing less than n % of cases. To reduce table size, you can omit statistics that are computed for only a small number of cases.

To Display Descriptive Statistics

1. In the main Missing Value Analysis dialog box, select the variable(s) for which you want to display missing value descriptive statistics.
2. Click **Descriptives**.
3. Choose the descriptive statistics that you want to display.

Estimating Statistics and Imputing Missing Values

You can choose to estimate means, standard deviations, covariances, and correlations using listwise (complete cases only), pairwise, EM (expectation-maximization), and/or regression methods. You can also choose to impute the missing values (estimate replacement values). Note that Multiple Imputation is generally considered to be superior to single imputation for solving the problem of missing values. Little's MCAR test is still useful for determining whether imputation is necessary.

Listwise Method

This method uses only complete cases. If any of the analysis variables have missing values, the case is omitted from the computations.

Pairwise Method

This method looks at pairs of analysis variables and uses a case only if it has nonmissing values for both of the variables. Frequencies, means, and standard deviations are computed separately for each pair.

Because other missing values in the case are ignored, correlations and covariances for two variables do not depend on values missing in any other variables.

EM Method

This method assumes a distribution for the partially missing data and bases inferences on the likelihood under that distribution. Each iteration consists of an E step and an M step. The E step finds the conditional expectation of the "missing" data, given the observed values and current estimates of the parameters. These expectations are then substituted for the "missing" data. In the M step, maximum likelihood estimates of the parameters are computed as though the missing data had been filled in. "Missing" is enclosed in quotation marks because the missing values are not being directly filled in. Instead, functions of them are used in the log-likelihood.

Roderick J. A. Little's chi-square statistic for testing whether values are missing completely at random (MCAR) is printed as a footnote to the EM matrices. For this test, the null hypothesis is that the data are missing completely at random, and the p value is significant at the 0.05 level. If the value is less than 0.05, the data are not missing completely at random. The data may be missing at random (MAR) or not missing at random (NMAR). You cannot assume one or the other and need to analyze the data to determine how the data are missing.

Regression Method

This method computes multiple linear regression estimates and has options for augmenting the estimates with random components. To each predicted value, the procedure can add a residual from a randomly selected complete case, a random normal deviate, or a random deviate (scaled by the square root of the residual mean square) from the t distribution.

EM Estimation Options

Using an iterative process, the EM method estimates the means, the covariance matrix, and the correlation of quantitative (scale) variables with missing values.

Distribution. EM makes inferences based on the likelihood under the specified distribution. By default, a normal distribution is assumed. If you know that the tails of the distribution are longer than those of a normal distribution, you can request that the procedure constructs the likelihood function from a Student's t distribution with n degrees of freedom. The mixed normal distribution also provides a distribution with longer tails. Specify the ratio of the standard deviations of the mixed normal distribution and the mixture proportion of the two distributions. The mixed normal distribution assumes that only the standard deviations of the distributions differ. The means must be the same.

Maximum iterations. Sets the maximum number of iterations to estimate the true covariance. The procedure stops when this number of iterations is reached, even if the estimates have not converged.

Save completed data. You can save a dataset with the imputed values in place of the missing values. Be aware, though, that covariance-based statistics using the imputed values will underestimate their respective parameter values. The degree of underestimation is proportional to the number of cases that are jointly unobserved.

To Specify EM Options

1. In the main Missing Value Analysis dialog box, select the variable(s) for which you want to estimate missing values using the EM method.
2. Select **EM** in the Estimation group.
3. To specify predicted and predictor variables, click **Variables**. See the topic "Predicted and Predictor Variables" on page 6 for more information.
4. Click **EM**.
5. Select the EM options you want.

Regression Estimation Options

The regression method estimates missing values using multiple linear regression. The means, the covariance matrix, and the correlation matrix of the predicted variables are displayed.

Estimation Adjustment. The regression method can add a random component to regression estimates. You can select residuals, normal variates, Student's *t* variates, or no adjustment.

- *Residuals*. Error terms are chosen randomly from the observed residuals of complete cases to be added to the regression estimates.
- *Normal Variates*. Error terms are randomly drawn from a distribution with the expected value 0 and the standard deviation equal to the square root of the mean squared error term of the regression.
- *Student's t Variates*. Error terms are randomly drawn from a t distribution with the specified degrees of freedom, and scaled by the root mean squared error (RMSE).

Maximum number of predictors. Sets a maximum limit on the number of predictor (independent) variables used in the estimation process.

Save completed data. Writes a dataset in the current session or an external IBM SPSS Statistics data file, with missing values replaced by values estimated by the regression method.

To Specify Regression Options

1. In the main Missing Value Analysis dialog box, select the variable(s) for which you want to estimate missing values using the regression method.
2. Select **Regression** in the Estimation group.
3. To specify predicted and predictor variables, click **Variables**. See the topic "Predicted and Predictor Variables" on page 6 for more information.
4. Click **Regression**.
5. Select the regression options you want.

Predicted and Predictor Variables

By default, all quantitative variables are used for EM and regression estimation. If needed, you can choose specific variables as predicted and predictor variables in the estimation(s). A given variable can be in both lists, but there are situations in which you might want to restrict the use of a variable. For example, some analysts are uncomfortable estimating values of outcome variables. You may also want to use different variables for different estimations and run the procedure multiple times. For example, if you have a set of items that are nurses' ratings and another set that are doctors' ratings, you may want to make one run using the nurses' item to estimate missing nurses' items and another run for estimates of the doctors' items.

Another consideration arises when using the regression method. In multiple regression, the use of a large subset of independent variables can produce poorer predicted values than a smaller subset. Therefore, a variable must achieve an *F*-to-enter limit of 4.0 to be used. This limit can be changed with syntax.

To Specify Predicted and Predictor Variables

1. In the main Missing Value Analysis dialog box, select the variable(s) for which you want to estimate missing values using the regression method.
2. Select **EM** or **Regression** in the Estimation group.
3. Click **Variables**.
4. If you want to use specific rather than all variables as predicted and predictor variables, select **Select variables** and move variables to the appropriate list(s).

MVA Command Additional Features

The command syntax language also allows you to:

- Specify separate descriptive variables for missing value patterns, data patterns, and tabulated patterns using the DESCRIBE keyword on the MPATTERN, DPATTERN, or TPATTERN subcommands.
- Specify more than one sort variable for the data patterns table, using the DPATTERN subcommand.
- Specify more than one sort variable for data patterns, using the DPATTERN subcommand.

- Specify tolerance and convergence, using the EM subcommand.
- Specify tolerance and *F*-to-enter, using the REGRESSION subcommand.
- Specify different variable lists for EM and Regression, using the EM and REGRESSION subcommands.
- Specify different percentages for suppressing cases displayed, for each of TTESTS, TABULATE, and MISMATCH.

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

Multiple Imputation

The purpose of multiple imputation is to generate possible values for missing values, thus creating several "complete" sets of data. Analytic procedures that work with multiple imputation datasets produce output for each "complete" dataset, plus pooled output that estimates what the results would have been if the original dataset had no missing values. These pooled results are generally more accurate than those provided by single imputation methods.

Multiple Imputation Data Considerations

Analysis variables. The analysis variables can be:

- *Nominal*. A variable can be treated as nominal when its values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, postal code, and religious affiliation.
- *Ordinal*. A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.
- *Scale*. A variable can be treated as scale (continuous) when its values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variables include age in years and income in thousands of dollars.

  The procedure assumes that the appropriate measurement level has been assigned to all variables; however, you can temporarily change the measurement level for a variable by right-clicking the variable in the source variable list and selecting a measurement level from the pop-up menu. To permanently change the level of measurement for a variable,

An icon next to each variable in the variable list identifies the measurement level and data type:

Table 1. Measurement level icons				
	Numeric	**String**	**Date**	**Time**
Scale (Continuous)		n/a		
Ordinal				
Nominal				

Frequency weights. Frequency (replication) weights are honored by this procedure. Cases with negative or zero replication weight value are ignored. Noninteger weights are rounded to the nearest integer.

Analysis Weight. Analysis (regression or sampling) weights are incorporated in summaries of missing values and in fitting imputation models. Cases with a negative or zero analysis weight are excluded.

Complex Samples. The Multiple Imputation procedure does not explicitly handle strata, clusters, or other complex sampling structures, though it can accept final sampling weights in the form of the analysis weight variable. Also note that Complex Sampling procedures currently do not automatically analyze

multiply imputed datasets. For a full list of procedures that support pooling, see "Analyzing Multiple Imputation Data" on page 13.

Missing Values. Both user- and system-missing values are treated as invalid values; that is, both types of missing values are replaced when values are imputed and both are treated as invalid values of variables used as predictors in imputation models. User- and system-missing values are also treated as missing in analyses of missing values.

Replicating results (Impute Missing Data Values). If you want to replicate your imputation results exactly, use the same initialization value for the random number generator, the same data order, and the same variable order, in addition to using the same procedure settings.

- **Random number generation.** The procedure uses random number generation during calculation of imputed values. To reproduce the same randomized results in the future, use the same initialization value for the random number generator before each run of the Impute Missing Data Values procedure.
- **Case order.** Values are imputed in case order.
- **Variable order.** The fully conditional specification (FCS) imputation method imputes values in the order specified in the Analysis Variables list.

There are two dialogs dedicated to multiple imputation.

- Analyze Patterns provides descriptive measures of the patterns of missing values in the data, and can be useful as an exploratory step before imputation.
- Impute Missing Data Values is used to generate multiple imputations. The complete datasets can be analyzed with procedures that support multiple imputation datasets. See Random Number Generators for information on analyzing multiple imputation datasets and a list of procedures that support these data.

Related information

Analyze Patterns

Analyze Patterns provides descriptive measures of the patterns of missing values in the data, and can be useful as an exploratory step before imputation.

Example. A telecommunications provider wants to better understand service usage patterns in its customer database. They have complete data for services used by their customers, but the demographic information collected by the company has a number of missing values. Analyzing the patterns of missing values can help determine next steps for imputation.

From the menus choose:

Analyze > **Multiple Imputation** > **Analyze Patterns...**

1. Select at least two analysis variables. The procedure analyzes patterns of missing data for these variables.

Optional Settings

Analysis Weight. This variable contains analysis (regression or sampling) weights. The procedure incorporates analysis weights in summaries of missing values. Cases with a negative or zero analysis weight are excluded.

Output. The following optional output is available:

- **Summary of missing values.** This displays a paneled pie chart that shows the number and percent of analysis variables, cases, or individual data values that have one or more missing values.
- **Patterns of missing values.** This displays tabulated patterns of missing values. Each pattern corresponds to a group of cases with the same pattern of incomplete and complete data on analysis variables. You can use this output to determine whether the monotone imputation method can be used for your data, or if not, how closely your data approximate a monotone pattern. The procedure orders analysis variables to reveal or approximate a monotonic pattern. If no nonmonotone pattern exists

after reordering you can conclude that the data have a monotonic pattern when analysis variables are ordered as such.

- **Variables with the highest frequency of missing values.** This displays a table of analysis variables sorted by percent of missing values in decreasing order. The table includes descriptive statistics (mean and standard deviation) for scale variables.

  You can control the maximum number of variables to display and minimum percentage missing for a variable to be included in the display. The set of variables that meet both criteria are displayed. For example, setting the maximum number of variables to 50 and the minimum percentage missing to 25 requests that the table display up to 50 variables that have at least 25% missing values. If there are 60 analysis variables but only 15 have 25% or more missing values, the output includes only 15 variables.

Impute Missing Data Values

Impute Missing Data Values is used to generate multiple imputations. The complete datasets can be analyzed with procedures that support multiple imputation datasets. See "Analyzing Multiple Imputation Data" on page 13 for information on analyzing multiple imputation datasets and a list of procedures that support these data.

Example. A telecommunications provider wants to better understand service usage patterns in its customer database. They have complete data for services used by their customers, but the demographic information collected by the company has a number of missing values. Moreover, these values are not missing completely at random, so multiple imputation will be used to complete the dataset.

From the menus choose:

Analyze > **Multiple Imputation** > **Impute Missing Data Values...**

1. Select at least two variables in the imputation model. The procedure imputes multiple values for missing data for these variables.
2. Specify the number of imputations to compute. By default, this value is 5.
3. Specify a dataset or IBM SPSS Statistics-format data file to which imputed data should be written.

   The output dataset consists of the original case data with missing data plus a set of cases with imputed values for each imputation. For example, if the original dataset has 100 cases and you have five imputations, the output dataset will have 600 cases. All variables in the input dataset are included in the output dataset. Dictionary properties (names, labels, etc.) of existing variables are copied to the new dataset. The file also contains a new variable, *Imputation_*, a numeric variable that indicates the imputation (0 for original data, or 1..*n* for cases having imputed values).

   The procedure automatically defines the *Imputation_* variable as a split variable when the output dataset is created. If splits are in effect when the procedure executes, the output dataset includes one set of imputations for each combination of values of split variables.

Optional Settings

Analysis Weight. This variable contains analysis (regression or sampling) weights. The procedure incorporates analysis weights in regression and classification models used to impute missing values. Analysis weights are also used in summaries of imputed values; for example, mean, standard deviation, and standard error. Cases with a negative or zero analysis weight are excluded.

Fields with unknown measurement level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Scan Data. Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.

Assign Manually. Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

Method

The Method tab specifies how missing values will be imputed, including the types of models used. Categorical predictors are indicator (dummy) coded.

Imputation Method

The **Automatic** method scans the data and uses the monotone method if the data show a monotone pattern of missing values; otherwise, fully conditional specification is used. If you are certain of which method you want to use, you can specify it as a **Custom** method.

Fully conditional specification

This is an iterative Markov chain Monte Carlo (MCMC) method that can be used when the pattern of missing data is arbitrary (monotone or non-monotone).

For each iteration and for each variable in the order specified in the variable list, the fully conditional specification (FCS) method fits a univariate (single dependent variable) model using all other available variables in the model as predictors, then imputes missing values for the variable being fit. The method continues until the maximum number of iterations is reached, and the imputed values at the maximum iteration are saved to the imputed dataset.

Maximum iterations

This specifies the number of iterations, or "steps", taken by the Markov chain used by the FCS method. If the FCS method was chosen automatically, it uses the default number of 10 iterations. When you explicitly choose FCS, you can specify a custom number of iterations. You may need to increase the number of iterations if the Markov chain hasn't converged. On the Output tab, you can save FCS iteration history data and plot it to assess convergence.

Monotone

This is a non-iterative method that can be used only when the data have a monotone pattern of missing values. A monotone pattern exists when you can order the variables such that, if a variable has a non-missing value, all preceding variables also have non-missing values. When specifying this as a **Custom** method, be sure to specify the variables in the list in an order that shows a monotone pattern.

For each variable in the monotone order, the monotone method fits a univariate (single dependent variable) model using all preceding variables in the model as predictors, then imputes missing values for the variable being fit. These imputed values are saved to the imputed dataset.

Include two-way interactions

When the imputation method is chosen automatically, the imputation model for each variable includes a constant term and main effects for predictor variables. When choosing a specific method, you can optionally include all possible two-way interactions among categorical predictor variables.

Model type for scale variables

Linear Regression

When the imputation method is chosen automatically, linear regression is used as the univariate model for scale variables.

Predictive Mean Matching (PMM)

When choosing a specific method, you can alternatively choose predictive mean matching (PMM) as the model for scale variables. PMM is a variant of linear regression that ensures that the imputed values are plausible. For PMM, the imputed value is based on the value defined for the **Randomly select a complete case from the closest (k) predictions** value, where (k) is a positive integer with a default value of 5.

Logistic regression is always used as the univariate model for categorical variables. Regardless of the model type, categorical predictors are handled using indicator (dummy) coding.

Singularity tolerance

Singular (or non-invertible) matrices have linearly dependent columns, which can cause serious problems for the estimation algorithm. Even near-singular matrices can lead to poor results, so the

procedure will treat a matrix whose determinant is less than the tolerance as singular. Specify a positive value.

Constraints

The Constraints tab allows you to restrict the role of a variable during imputation and restrict the range of imputed values of a scale variable so that they are plausible. In addition, you can restrict the analysis to variables with less than a maximum percentage of missing values.

Scan of Data for Variable Summary. Clicking **Scan Data** causes the list to show analysis variables and the observed percent missing, minimum, and maximum for each. The summaries can be based on all cases or limited to a scan of the first *n* cases, as specified in the Cases text box. Clicking **Rescan Data** updates the distribution summaries.

Define Constraints

- **Role.** This allows you to customize the set of variables to be imputed and/or treated as predictors. Typically, each analysis variable is considered as both a dependent and predictor in the imputation model. The **Role** can be used to turn off imputation for variables that you want to **Use as predictor only** or to exclude variables from being used as predictors (**Impute only**) and thereby make the prediction model more compact. This is the only constraint that may be specified for categorical variables, or for variables that are used as predictors only.
- **Min and Max.** These columns allow you to specify minimum and maximum allowable imputed values for scale variables. If an imputed value falls outside this range, the procedure draws another value until it finds one within the range or the maximum number of draws is reached (see **Maximum draws** below). These columns are only available if **Linear Regression** is selected as the scale variable model type on the Method tab.
- **Rounding.** Some variables may be used as scale, but have values that are naturally further restricted; for instance, the number of people in a household must be integer, and the amount spent during a visit to the grocery store cannot have fractional cents. This column allows you to specify the smallest denomination to accept. For example, to obtain integer values you would specify 1 as the rounding denomination; to obtain values rounded to the nearest cent, you would specify 0.01. In general, values are rounded to the nearest integer multiple of the rounding denomination. The following table shows how different rounding values act upon an imputed value of 6.64823 (before rounding).

Table 2. Rounding results	
Rounding Denomination	**Value to which 6.64832 is rounded**
10	10
1	7
0.25	6.75
0.1	6.6
0.01	6.65

Exclude variables with large amounts of missing data. Typically, analysis variables are imputed and used as predictors without regard to how many missing values they have, provided they have sufficient data to estimate an imputation model. You can choose to exclude variables that have a high percentage of missing values. For example, if you specify 50 as the **Maximum percentage missing**, analysis variables that have more than 50% missing values are not imputed, nor are they used as predictors in imputation models.

Maximum draws. If minimum or maximum values are specified for imputed values of scale variables (see **Min and Max** above), the procedure attempts to draw values for a case until it finds a set of values that are within the specified ranges. If a set of values is not obtained within the specified number of draws per case, the procedure draws another set of model parameters and repeats the case-drawing process. An error occurs if a set of values within the ranges is not obtained within the specified number of case and parameter draws.

Note that increasing these values can increase the processing time. If the procedure is taking a long time, or is unable to find suitable draws, check the minimum and maximum values specified to ensure they are appropriate.

Output

Display. Controls display of output. An overall imputation summary is always displayed, which includes tables relating the imputation specifications, iterations (for fully conditional specification method), dependent variables imputed, dependent variables excluded from imputation, and imputation sequence. If specified, constaints for analysis variables are also shown.

- **Imputation model.** This displays the imputation model for dependent variables and predictors, and includes univariate model type, model effects, and number of values imputed.
- **Descriptive statistics.** This displays descriptive statistics for dependent variables for which values are imputed. For scale variables the descriptive statistics include mean, count, standard deviation, min, and max for the original input data (prior to imputation), imputed values (by imputation), and complete data (original and imputed values together—by imputation). For categorical variables the descriptive statistics include count and percent by category for the original input data (prior to imputation), imputed values (by imputation), and complete data (original and imputed values together—by imputation).

Iteration History. When the fully conditional specification imputation method is used, you can request a dataset that contains iteration history data for FCS imputation. The dataset contains means and standard deviations by iteration and imputation for each scale dependent varable for which values are imputed. You can plot the data to help assess model convergence.

MULTIPLE IMPUTATION Command Additional Features

The command syntax language also allows you to:

- Specify a subset of variables for which descriptive statistics are shown (IMPUTATIONSUMMARIES subcommand).
- Specify both an analysis of missing patterns and imputation in a single run of the procedure.
- Specify the maximum number of model parameters allowed when imputing any variable (MAXMODELPARAM keyword).

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

Working with Multiple Imputation Data

When a multiple imputation (MI) dataset is created, a variable called *Imputation_*, with variable label *Imputation Number*, is added, and the dataset is sorted by it in ascending order. Cases from the original dataset has a value of 0. Cases for imputed values are numbered 1 through *M*, where *M* is the number of imputations.

When you open a dataset, the presence of *Imputation_* identifies the dataset as a possible MI dataset.

Activating a Multiple Imputation Dataset for Analysis

The dataset must be split using the **Compare groups** option, with *Imputation_* as a grouping variable, in order to be treated as an MI dataset in analyses. You can also define splits on other variables.

From the menus choose:

Data > **Split File...**

1. Select **Compare groups**.
2. Select *Imputation Number [Imputation_]* as a variable to group cases on.

Alternatively, when you turn markings on (see below), the the file is split on *Imputation Number [Imputation_]*.

Distinguishing Imputed Values from Observed Values

You can distinguish imputed values from observed values by cell background color, the font, and bold type (for imputed values). When you create a new dataset in the current session with Impute Missing Values, markings are turned on by default. When you open a saved data file that includes imputations, markings are turned off.

To turn markings on, from the Data Editor menus choose:

View > **Mark Imputed Data...**

Alternatively, you can turn on markings by clicking the imputation marking button at the right edge of the edit bar in Data View of the Data Editor.

Moving Between Imputations

1. From the menus choose:

   Edit > **Go to Imputation...**
2. Select the imputation (or Original data) from the drop-down list.

Alternatively, you can select the imputation from the drop-down list in the edit bar in Data View of the Data Editor.

Relative case position is preserved when selecting imputations. For example, if there are 1000 cases in the original dataset, case 1034, the 34th case in the first imputation, displays at the top of the grid. If you select imputation **2** in the dropdown, case 2034, the 34th case in imputation 2, would display at the top of the grid. If you select **Original data** in the dropdown, case 34 would display at the top of the grid. Column position is also preserved when navigating between imputations, so that it is easy to compare values between imputations.

Transforming and Editing Imputed Values

Sometimes you will need to perform transformations on imputed data. For example, you may want to take the log of all values of a salary variable and save the result in a new variable. A value computed using imputed data will be treated as imputed if it differs from the value computed using the original data.

If you edit an imputed value in a cell of the Data Editor, that cell is still treated as imputed. It is not recommended to edit imputed values in this way.

Analyzing Multiple Imputation Data

Many procedures support pooling of results from analysis of multiply imputed data sets. When imputation markings are turned on, a special icon is displayed next to procedures that support pooling. On the Descriptive Statistics submenu of the Analyze menu, for example, Frequencies, Descriptives, Explore, and Crosstabs all support pooling, while Ratio, P-P Plots, and Q-Q Plots do not.

Both tabular output and model PMML can be pooled. There is no new procedure for requesting pooled output; instead, a new tab on the Options dialog gives you global control over multiple imputation outputs.

- **Pooling of Tabular Output.** By default, when you run a supported procedure on a multiple imputation (MI) data set, results are automatically produced for each imputation, the original (unimputed) data, and pooled (final) results that take into account variation across imputations. The statistics that are pooled vary by procedure.

- **Pooling of PMML.** You can also obtain pooled PMML from supported procedures that export PMML. Pooled PMML is requested in the same way as, and is saved instead of, non pooled PMML.

Unsupported procedures produce neither pooled output nor pooled PMML files.

Levels of Pooling

Output is pooled by using one of two levels:

- **Naïve combination.** Only the pooled parameter is available.

- **Univariate combination.** The pooled parameter, its standard error, test statistic and effective degrees of freedom, *p*-value, confidence interval, and pooling diagnostics (fraction of missing information, relative efficiency, relative increase in variance) are shown when available.

Coefficients (regression and correlation), means (and mean differences), and counts are typically pooled. When the standard error of the statistic is available, then univariate pooling is used; otherwise, naïve pooling is used.

Procedures That Support Pooling

The following procedures support MI data sets, at the levels of pooling specified for each piece of output.

Frequencies. The following features are supported:

- The Statistics table supports Means at Univariate pooling (if S.E. mean is also requested) and Valid N and Missing N at Naïve pooling.
- The Frequencies table supports Frequency at Naïve pooling.

Descriptives. The following features are supported:

- The Descriptive Statistics table supports Means at Univariate pooling (if S.E. mean is also requested) and N at Naïve pooling.

Crosstabs. The following features are supported:

- The Crosstabulation table supports Count at Naïve pooling.

Means. The following features are supported:

- The Report table supports Mean at Univariate pooling (if S.E. mean is also requested) and N at Naïve pooling.

One-Sample T Test. The following features are supported:

- The Statistics table supports Mean at Univariate pooling and N at Naïve pooling.
- The Test table supports Mean Difference at Univariate pooling.

Independent-Samples T Test. The following features are supported:

- The Group Statistics table supports Means at Univariate pooling and N at Naïve pooling.
- The Test table supports Mean Difference at Univariate pooling.

Paired-Samples T Test. The following features are supported:

- The Statistics table supports Means at Univariate pooling and N at Naïve pooling.
- The Correlations table supports Correlations and N at Naïve pooling.
- The Test table supports Mean at Univariate pooling.

One-Way ANOVA. The following features are supported:

- The Descriptive Statistics table supports Mean at Univariate pooling and N at Naïve pooling.
- The Contrast Tests table supports Value of Contrast at Univariate pooling.

GLM Univariate and Multivariate. The following features are supported:

- The Between-Subjects Factors table supports N at Naive pooling.
- The Descriptive Statistics table supports Mean and N at Naive pooling.
- The Parameter Estimates table supports the coefficient, B, at Univariate pooling.
- The Estimated Marginal Means: Estimates table supports Mean at Univariate pooling.
- The Estimated Marginal Means: Pairwise Comparisons table supports Mean Difference at Univariate pooling.

Linear Mixed Models. The following features are supported:

- The Descriptive Statistics table supports Mean and N at Naïve pooling.

- The Estimates of Fixed Effects table supports Estimate at Univariate pooling.
- The Estimates of Covariance Parameters table supports Estimate at Univariate pooling.
- The Estimated Marginal Means: Estimates table supports Mean at Univariate pooling.
- The Estimated Marginal Means: Pairwise Comparisons table supports Mean Difference at Univariate pooling.

Generalized Linear Models and Generalized Estimating Equations. These procedures support pooled PMML.

- The Categorical Variable Information table supports N and Percents at Naïve pooling.
- The Continuous Variable Information table supports N and Mean at Naïve pooling.
- The Parameter Estimates table supports the coefficient, B, at Univariate pooling.
- The Estimated Marginal Means: Estimation Coefficients table supports Mean at Naïve pooling.
- The Estimated Marginal Means: Estimates table supports Mean at Univariate pooling.
- The Estimated Marginal Means: Pairwise Comparisons table supports Mean Difference at Univariate pooling.

Bivariate Correlations. The following features are supported:

- The Descriptive Statistics table supports Mean and N at Naïve pooling.
- The Correlations table supports Correlations and N at Univariate pooling. Note that correlations are transformed by using Fisher's *z* transformation before pooling, and then back transformed after pooling.

Partial Correlations. The following features are supported:

- The Descriptive Statistics table supports Mean and N at Naïve pooling.
- The Correlations table supports Correlations at Naïve pooling.

Linear Regression. This procedure supports pooled PMML.

- The Descriptive Statistics table supports Mean and N at Naïve pooling.
- The Correlations table supports Correlations and N at Naïve pooling.
- The Coefficients table supports B at Univariate pooling and Correlations at Naïve pooling.
- The Correlation Coefficients table supports Correlations at Naïve pooling.
- The Residuals Statistics table supports Mean and N at Naïve pooling.

Binary Logistic Regression. This procedure supports pooled PMML.

- The Variables in the Equation table supports B at Univariate pooling.

Multinomial Logistic Regression. This procedure supports pooled PMML.

- The Parameter Estimates table supports the coefficient, B, at Univariate pooling.

Ordinal Regression. The following features are supported:

- The Parameter Estimates table supports the coefficient, B, at Univariate pooling.

Discriminant Analysis. This procedure supports pooled model XML.

- The Group Statistics table supports Mean and Valid N at Naïve pooling.
- The Pooled Within-Groups Matrices table supports Correlations at Naïve pooling.
- The Canonical Discriminant Function Coefficients table supports Unstandardized Coefficients at Naïve pooling.
- The Functions at Group Centroids table supports Unstandardized Coefficients at Naïve pooling.
- The Classification Function Coefficients table supports Coefficients at Naïve pooling.

Chi-Square Test. The following features are supported:

- The Descriptives table supports Mean and N at Naïve pooling.

- The Frequencies table supports Observed N at Naïve pooling.

Binomial Test. The following features are supported:

- The Descriptives table supports Means and N at Naïve pooling.
- The Test table supports N, Observed Proportion, and Test Proportion at Naïve pooling.

Runs Test. The following features are supported:

- The Descriptives table supports Means and N at Naïve pooling.

One-Sample Kolmogorov-Smirnov Test. The following features are supported:

- The Descriptives table supports Means and N at Naïve pooling.

Two-Independent-Samples Tests. The following features are supported:

- The Ranks table supports Mean Rank and N at Naïve pooling.
- The Frequencies table supports N at Naïve pooling.

Tests for Several Independent Samples. The following features are supported:

- The Ranks table supports Mean Rank and N at Naïve pooling.
- The Frequencies table supports Counts at Naïve pooling.

Two-Related-Samples Tests. The following features are supported:

- The Ranks table supports Mean Rank and N at Naïve pooling.
- The Frequencies table supports N at Naïve pooling.

Tests for Several Related Samples. The following features are supported:

- The Ranks table supports Mean Rank at Naïve pooling.

Cox Regression. This procedure supports pooled PMML.

- The Variables in the Equation table supports B at Univariate pooling.
- The Covariate Means table supports Mean at Naïve pooling.

One-Sample Proportions. The following features are supported:

- The One-Sample Proportions Confidence Intervals table supports Proportion at Univariate pooling.
- The One-Sample Proportions Test table supports Observed - Test Value at Univariate pooling.

Independent-Samples Proportions. The following features are supported:

- The Independent-Samples Proportions Group Statistics table supports Proportion at Univariate pooling.
- The Independent-Samples Confidence Intervals table supports Difference in Proportions at Naïve pooling.
- The Independent-Samples Proportions Test table supports Difference in Proportions at Univariate pooling.

Paired-Samples Proportions. The following features are supported:

- The Paired-Samples Proportions Statistics table supports Proportion at Univariate pooling.
- The Paired-Samples Confidence Intervals table supports Difference in Proportions at Naïve pooling.
- The Paired-Samples Proportions Test table supports Difference in Proportions at Univariate pooling.

Multiple Imputation Options

The Multiple Imputations tab controls two kinds of preferences related to Multiple Imputations:

Marking of Imputed Data. By default, cells containing imputed data will have a different background color than cells containing nonimputed data. The distinctive appearance of the imputed data should make it easy for you to scroll through a dataset and locate those cells. You can change the default cell background color, the font, and make the imputed data display in bold type.

Analysis Output. This group controls the type of Viewer output produced whenever a multiply imputed dataset is analyzed. By default, output will be produced for the original (pre-imputation) dataset and for each of the imputed datasets. In addition, for those procedures that support pooling of imputed data, final pooled results will be generated. When univariate pooling is performed, pooling diagnostics will also display. However, you can suppress any output you do not want to see.

To Set Multiple Imputation Options

From the menus, choose:

Edit > **Options**

Click the Multiple Imputation tab.

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