

# > SPSS Missing Values™ 17.0



For more information about SPSS Inc. software products, please visit our Web site at <http://www.spss.com> or contact

SPSS Inc.

233 South Wacker Drive, 11th Floor

Chicago, IL 60606-6412

Tel: (312) 651-3000

Fax: (312) 651-3668

SPSS is a registered trademark and the other product names are the trademarks of SPSS Inc. for its proprietary computer software. No material describing such software may be produced or distributed without the written permission of the owners of the trademark and license rights in the software and the copyrights in the published materials.

The SOFTWARE and documentation are provided with RESTRICTED RIGHTS. Use, duplication, or disclosure by the Government is subject to restrictions as set forth in subdivision (c) (1) (ii) of The Rights in Technical Data and Computer Software clause at 52.227-7013. Contractor/manufacturer is SPSS Inc., 233 South Wacker Drive, 11th Floor, Chicago, IL 60606-6412.

Patent No. 7,023,453

General notice: Other product names mentioned herein are used for identification purposes only and may be trademarks of their respective companies.

Windows is a registered trademark of Microsoft Corporation.

Apple, Mac, and the Mac logo are trademarks of Apple Computer, Inc., registered in the U.S. and other countries.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, <http://www.winwrap.com>.

Printed in the United States of America.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

---

# ***Preface***

SPSS Statistics 17.0 is a comprehensive system for analyzing data. The Missing Values optional add-on module provides the additional analytic techniques described in this manual. The Missing Values add-on module must be used with the SPSS Statistics 17.0 Base system and is completely integrated into that system.

## ***Installation***

To install the Missing Values add-on module, run the License Authorization Wizard using the authorization code that you received from SPSS Inc. For more information, see the installation instructions supplied with the Missing Values add-on module.

## ***Compatibility***

SPSS Statistics is designed to run on many computer systems. See the installation instructions that came with your system for specific information on minimum and recommended requirements.

## ***Serial Numbers***

Your serial number is your identification number with SPSS Inc. You will need this serial number when you contact SPSS Inc. for information regarding support, payment, or an upgraded system. The serial number was provided with your Base system.

## ***Customer Service***

If you have any questions concerning your shipment or account, contact your local office, listed on the Web site at <http://www.spss.com/worldwide>. Please have your serial number ready for identification.

### ***Training Seminars***

SPSS Inc. provides both public and onsite training seminars. All seminars feature hands-on workshops. Seminars will be offered in major cities on a regular basis. For more information on these seminars, contact your local office, listed on the Web site at <http://www.spss.com/worldwide>.

### ***Technical Support***

Technical Support services are available to maintenance customers. Customers may contact Technical Support for assistance in using SPSS Statistics or for installation help for one of the supported hardware environments. To reach Technical Support, see the Web site at <http://www.spss.com>, or contact your local office, listed on the Web site at <http://www.spss.com/worldwide>. Be prepared to identify yourself, your organization, and the serial number of your system.

### ***Additional Publications***

The *SPSS Statistical Procedures Companion*, by Marija Norušis, has been published by Prentice Hall. A new version of this book, updated for SPSS Statistics 17.0, is planned. The *SPSS Advanced Statistical Procedures Companion*, also based on SPSS Statistics 17.0, is forthcoming. The *SPSS Guide to Data Analysis* for SPSS Statistics 17.0 is also in development. Announcements of publications available exclusively through Prentice Hall will be available on the Web site at <http://www.spss.com/estore> (select your home country, and then click Books).

---

# Contents

## **Part I: User's Guide**

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b><i>Introduction to Missing Values</i></b>                   | <b>1</b>  |
| <b>2</b> | <b><i>Missing Value Analysis</i></b>                           | <b>3</b>  |
|          | Displaying Patterns of Missing Values . . . . .                | 6         |
|          | Displaying Descriptive Statistics for Missing Values . . . . . | 8         |
|          | Estimating Statistics and Imputing Missing Values . . . . .    | 10        |
|          | EM Estimation Options . . . . .                                | 11        |
|          | Regression Estimation Options . . . . .                        | 13        |
|          | Predicted and Predictor Variables . . . . .                    | 15        |
|          | MVA Command Additional Features . . . . .                      | 16        |
| <b>3</b> | <b><i>Multiple Imputation</i></b>                              | <b>17</b> |
|          | Analyze Patterns . . . . .                                     | 19        |
|          | Impute Missing Data Values . . . . .                           | 21        |
|          | Method . . . . .   | 24        |
|          | Constraints . . . . .  | 27        |
|          | Output . . . . .   | 30        |
|          | MULTIPLE IMPUTATION Command Additional Features . . . . .      | 31        |
|          | Working with Multiple Imputation Data . . . . .                | 31        |
|          | Analyzing Multiple Imputation Data . . . . .                   | 36        |
|          | Multiple Imputation Options . . . . .                          | 42        |

## ***Part II: Examples***

### **4 *Missing Value Analysis* 45**

|  |    |
|--|----|
| Describing the Pattern of Missing Data . . . . .                 | 45 |
| Running the Analysis to Display Descriptive Statistics . . . . . | 45 |
| Evaluating the Descriptive Statistics . . . . .                  | 47 |
| Rerunning the Analysis to Display Patterns . . . . .             | 54 |
| Evaluating the Patterns Table . . . . .                          | 56 |
| Rerunning the Analysis for Little's MCAR Test . . . . .          | 57 |

### **5 *Multiple Imputation* 59**

|   |    |
|---|----|
| Using Multiple Imputation to Complete and Analyze a Dataset . . . . . | 59 |
| Analyze Patterns of Missing Values . . . . .                          | 59 |
| Automatic Imputation of Missing Values . . . . .                      | 65 |
| Custom Imputation Model . . . . .                                     | 72 |
| Checking FCS Convergence . . . . .                                    | 82 |
| Analyze Complete Data . . . . .                                       | 87 |
| Summary . . . . .   | 99 |

## ***Appendix***

### **A *Sample Files* 100**

### ***Index* 114**

***Part I:***  
***User's Guide***





# *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:

- ▶ **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.

- ▶ **Impute missing values.** Use Impute Missing Data Values to multiply impute missing values.
- ▶ **Analyze “complete” data.** Use any procedure that supports multiple imputation data. See [Analyzing Multiple Imputation Data](#) on p. 36 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.

**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.

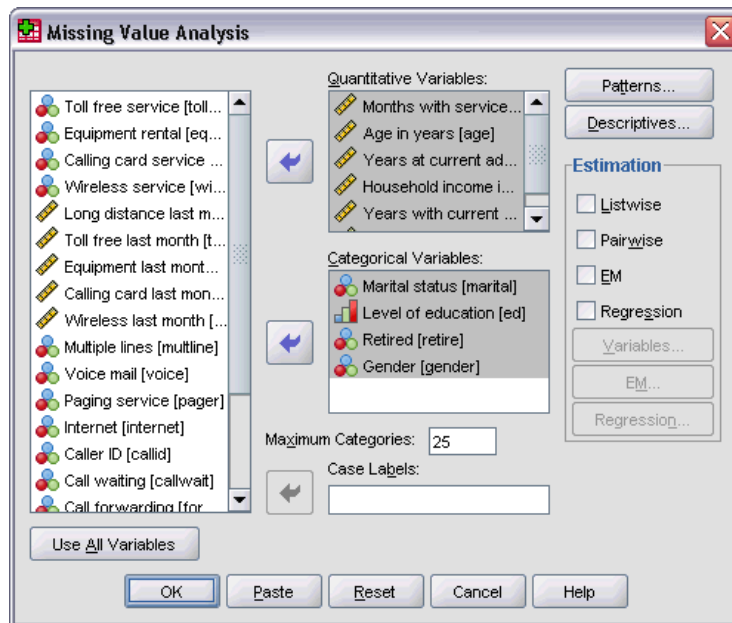
EM estimation depends on the assumption that the pattern of missing data is related to the observed data only. (This condition is called **missing at random**, or MAR.) This assumption allows estimates to be adjusted using available information. For example, in a study of education and income, the subjects 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 Trends add-on module, several methods are available to replace missing values in time series.

### To Obtain Missing Value Analysis

- ▶ From the menus choose:  
Analyze  
Missing Value Analysis...

Figure 2-1  
Missing Value Analysis dialog box



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

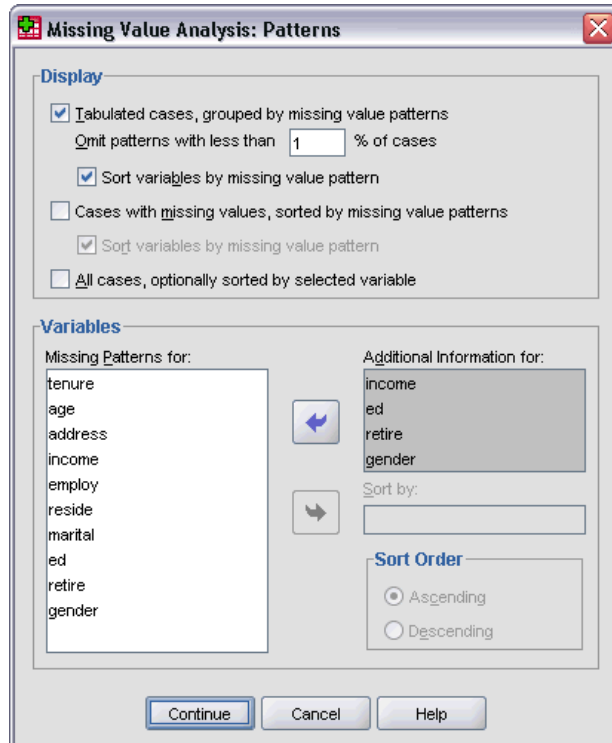
Optionally, you can:

- 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. For more information, see [Displaying Patterns of Missing Values](#) on p. 6.
- Click Descriptives to display descriptive statistics of missing values. For more information, see [Displaying Descriptive Statistics for Missing Values](#) on p. 8.

- Select a method for estimating statistics (means, covariances, and correlations) and possibly imputing missing values. For more information, see [Estimating Statistics and Imputing Missing Values](#) on p. 10.
- If you select EM or Regression, click Variables to specify a subset to be used for the estimation. For more information, see [Predicted and Predictor Variables](#) on p. 15.
- Select a case label variable. This variable is used to label cases in patterns tables that display individual cases.

## Displaying Patterns of Missing Values

Figure 2-2  
Missing Value Analysis Patterns dialog box



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.

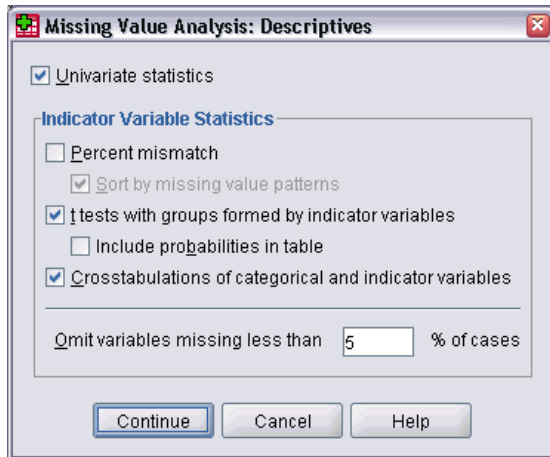
- **Sort by.** Cases are listed according to the ascending or descending order of the values of the specified variable. Available only for All cases.

### ***To Display Missing Value Patterns***

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

## ***Displaying Descriptive Statistics for Missing Values***

Figure 2-3  
*Missing Value Analysis Descriptives dialog box*



### ***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***

- ▶ In the main Missing Value Analysis dialog box, select the variable(s) for which you want to display missing value descriptive statistics.
- ▶ Click Descriptives.
- ▶ 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

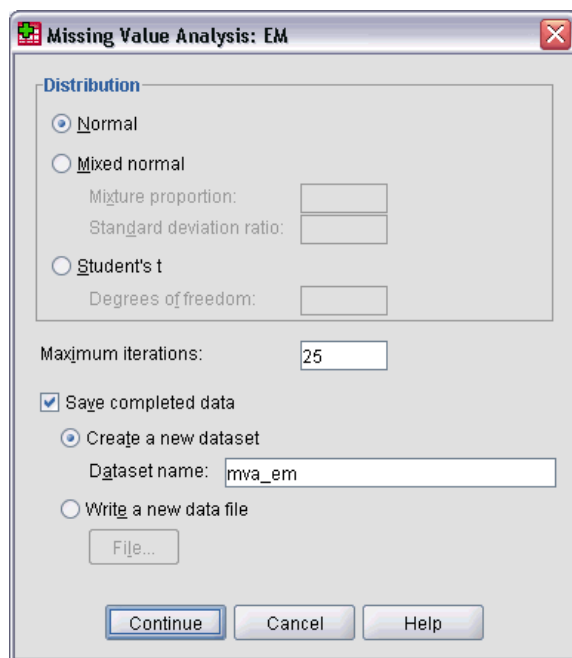
missing at random (NMAR). You cannot assume one or the other and need to analyze the data to determine how the data are missing.

### ***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***

Figure 2-4  
Missing Value Analysis EM dialog box



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.

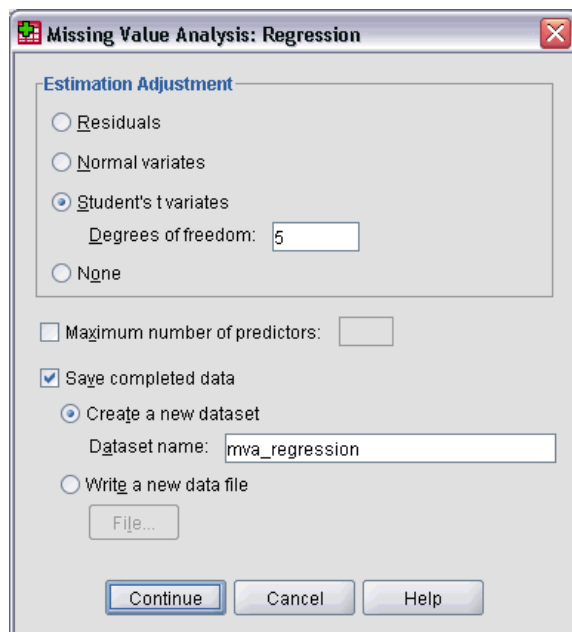
**Save completed data.** You can save a dataset with the imputed values in place of the missing values. Be aware, though, that covariance-based statistics using the imputed values will underestimate their respective parameter values. The degree of underestimation is proportional to the number of cases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

### ***To Specify EM Options***

- ▶ In the main Missing Value Analysis dialog box, select the variable(s) for which you want to estimate missing values using the EM method.
- ▶ Select EM in the Estimation group.
- ▶ To specify predicted and predictor variables, click Variables. For more information, see [Predicted and Predictor Variables](#) on p. 15.
- ▶ Click EM.
- ▶ Select the desired EM options.

## Regression Estimation Options

Figure 2-5  
Missing Value Analysis Regression dialog box



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.

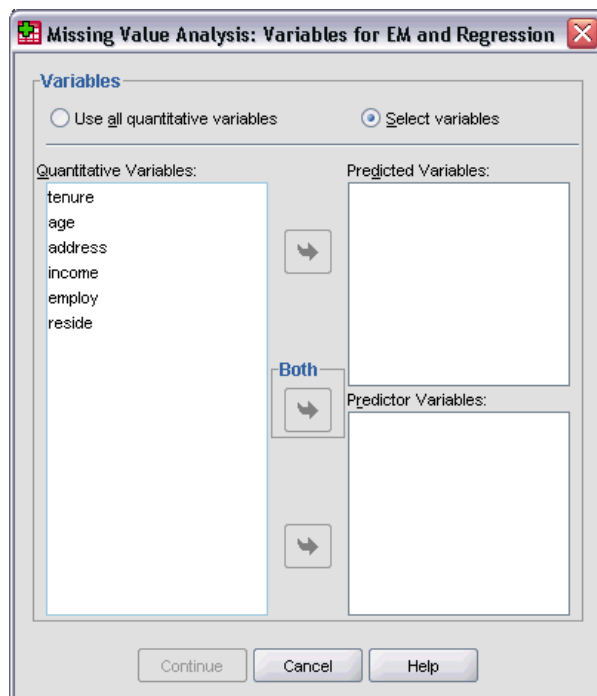
**Save completed data.** Writes a dataset in the current session or an external SPSS Statistics data file, with missing values replaced by values estimated by the regression method.

### ***To Specify Regression Options***

- ▶ In the main Missing Value Analysis dialog box, select the variable(s) for which you want to estimate missing values using the regression method.
- ▶ Select Regression in the Estimation group.
- ▶ To specify predicted and predictor variables, click Variables. For more information, see [Predicted and Predictor Variables](#) on p. 15.
- ▶ Click Regression.
- ▶ Select the desired regression options.

## Predicted and Predictor Variables

Figure 2-6  
Missing Value Analysis Variables for EM and Regression dialog box



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***

- ▶ In the main Missing Value Analysis dialog box, select the variable(s) for which you want to estimate missing values using the regression method.
- ▶ Select EM or Regression in the Estimation group.
- ▶ Click Variables.
- ▶ 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.

**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, zip 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 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 context menu.

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

| Measurement Level | Data Type   |   |   |   |
|-------------------|---|---|---|---|
|                   | Numeric   | String  | Date  | Time  |
| Scale             |  | 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 p. 36.

**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 [Analyzing Multiple Imputation Data](#) on p. 36 for information on analyzing multiple imputation datasets and a list of procedures that support these data.

## 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.

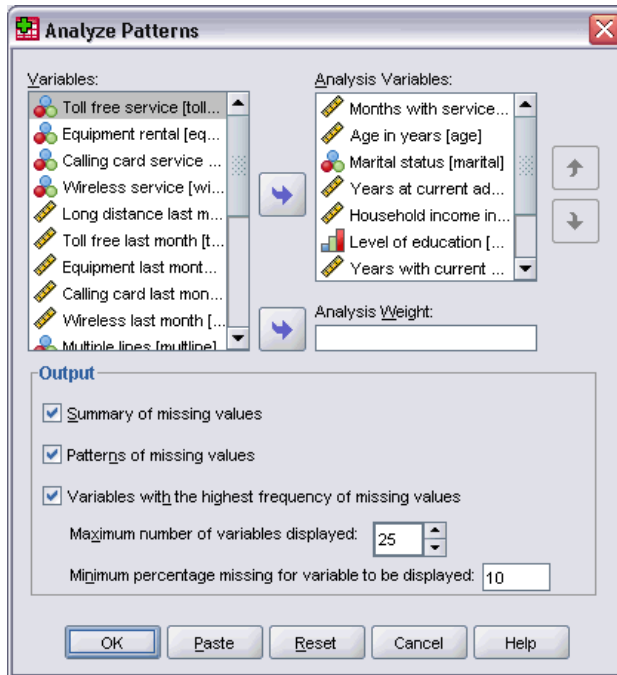
**Example.** A telecommunications provider wants to better understand service usage patterns in its customer database. They have complete data for services used by their customers, but the demographic information collected by the company has a number of missing values. Analyzing the patterns of missing values can help determine next steps for imputation. For more information, see [Using Multiple Imputation to Complete and Analyze a Dataset](#) in Chapter 5 on p. 59.

### *To Analyze Patterns of Missing Data*

From the menus choose:

```
Analyze
  Multiple Imputation
    Analyze Patterns...
```

Figure 3-1  
Analyze Patterns dialog box



- ▶ 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 p. 36 for information on analyzing multiple imputation datasets and a list of procedures that support these data.

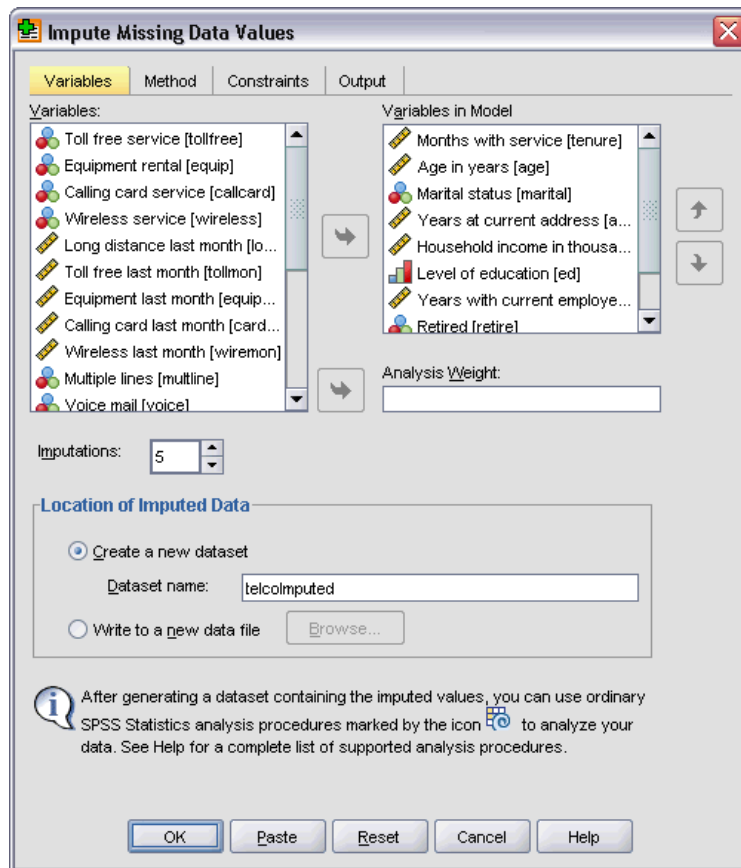
**Example.** A telecommunications provider wants to better understand service usage patterns in its customer database. They have complete data for services used by their customers, but the demographic information collected by the company has a number of missing values. Moreover, these values are not missing completely at random, so multiple imputation will be used to complete the dataset. For more information, see [Using Multiple Imputation to Complete and Analyze a Dataset](#) in Chapter 5 on p. 59.

### To Impute Missing Data Values

From the menus choose:

Analyze  
Multiple Imputation  
Impute Missing Data Values...

Figure 3-2  
*Impute Missing Data Values Variables tab*



- ▶ Select at least two variables in the imputation model. The procedure imputes multiple values for missing data for these variables.
- ▶ Specify the number of imputations to compute. By default, this value is 5.

- ▶ Specify a dataset or 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).

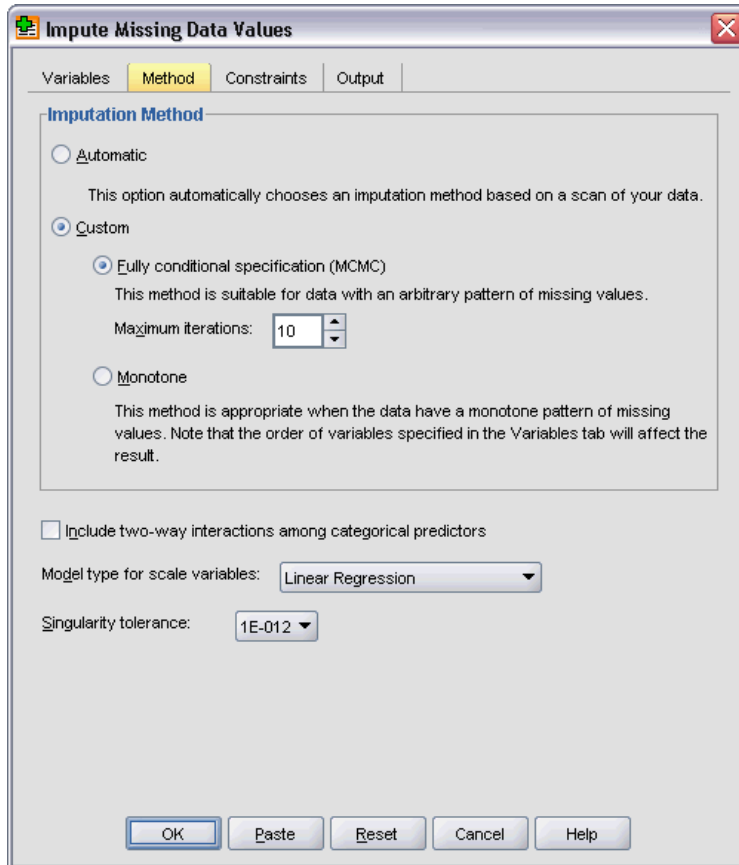
The procedure automatically defines the *Imputation\_* variable as a split variable when the output dataset is created. If splits are in effect when the procedure executes, the output dataset includes one set of imputations for each combination of values of split variables.

### ***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.

## Method

Figure 3-3  
*Impute Missing Data Values Method tab*



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 nonmonotone).

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 noniterative 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 nonmissing value, all preceding variables also have nonmissing 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.** When the imputation method is chosen automatically, linear regression is used as the univariate model for scale variables. 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 matches imputed values computed by the regression model to the closest observed value.

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

Figure 3-4  
Impute Missing Data Values Constraints tab

**Impute Missing Data Values**

Variables | Method | **Constraints** | Output

Scan of Data for Variable Summary

Rescan Data  Limit number of cases scanned Cases: 5000

Variable Summary:

| Variables in Model | Percent Missing | Observed Min | Observed Max |
|--------------------|-----------------|--------------|--------------|
| tenure             | 3.20            | 1            | 72           |
| age                | 2.50            | 18           | 77           |
| marital            | 11.50           | 0            | 1            |
| address            | 15.00           | 0            | 55           |

Cases Scanned: 1000

Define Constraints:

| Variables in Model | Role                        | Min | Max | Rounding |
|--------------------|-----------------------------|-----|-----|----------|
| retire             | impute and use as predictor |     |     |          |
| gender             | impute and use as predictor |     |     |          |
| reside             | impute and use as predictor | 1   |     | 1        |
| lninc              | impute and use as predictor | 0   |     |          |

Exclude variables with large amounts of missing data  
Maximum percentage missing:

Maximum case draws:

Maximum parameter draws:

Increasing the maximum parameter draws can significantly increase analysis time.

OK Paste Reset Cancel Help

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).

| Rounding Denomination | Value to which 6.64823 is rounded |
|-----------------------|-----------------------------------|
| 10                    | 10                                |
| 1                     | 7                                 |
| 0.25                  | 6.75                              |
| 0.1                   | 6.6                               |
| 0.01                  | 6.65                              |

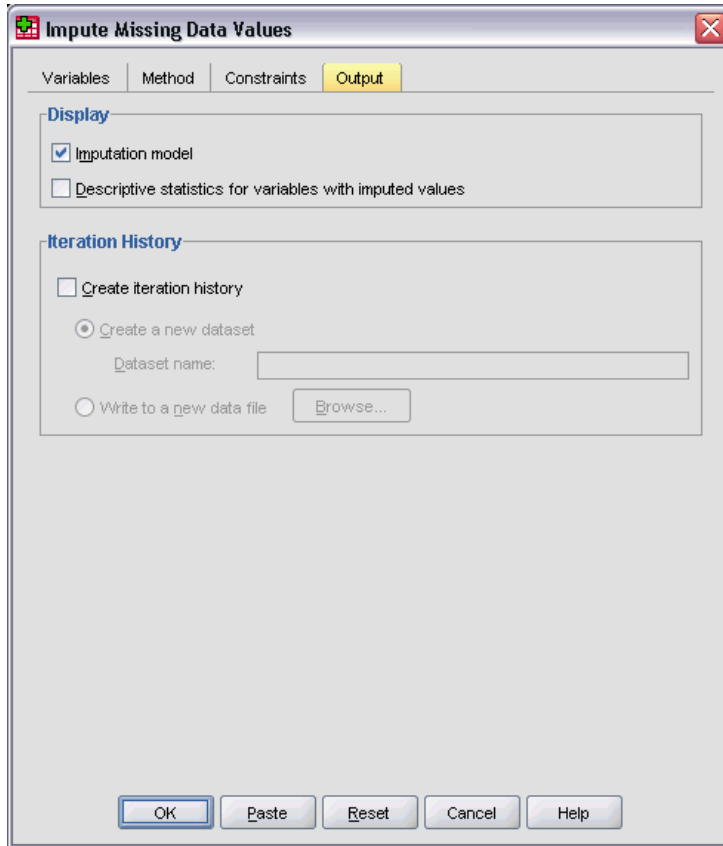
**Exclude variables with large amounts of missing data.** Typically, analysis variables are imputed and used as predictors without regard to how many missing values they have, provided they have sufficient data to estimate an imputation model. You can choose to exclude variables that have a high percentage 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

Figure 3-5  
*Impute Missing Data Values Output tab*



**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, constraints 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 variable for which values are imputed. You can plot the data to help assess model convergence. For more information, see [Checking FCS Convergence](#) in Chapter 5 on p. 82.

## ***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.

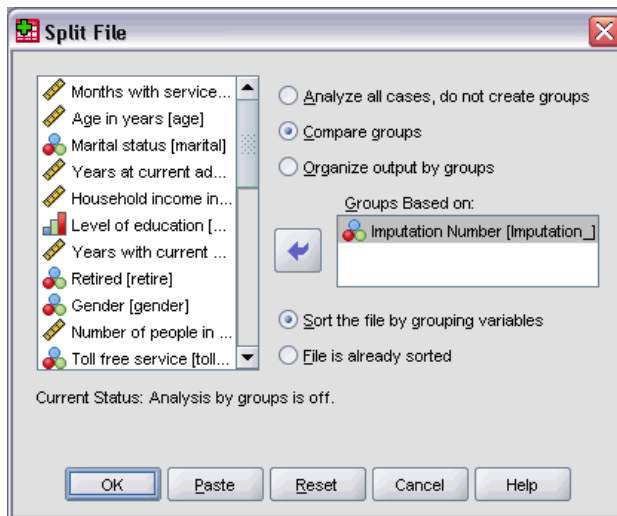
### **Activating a Multiple Imputation Dataset for Analysis**

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...

Figure 3-6  
*Split File dialog box*



- ▶ Select Compare groups.
- ▶ Select *Imputation Number [Imputation\_]* as a variable to group cases on.

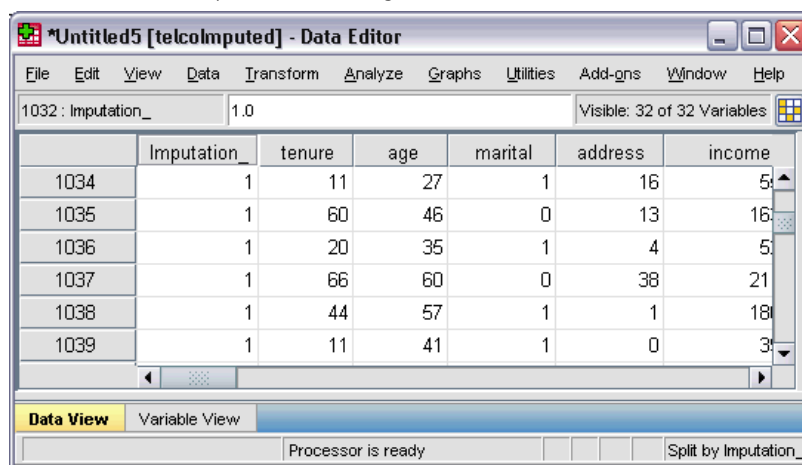
Alternatively, when you turn markings on (see below), the the file is split on *Imputation Number [Imputation\_]*.



### ***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). For details on which markings are in effect, see [Multiple Imputation Options](#) on p. 42. 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.

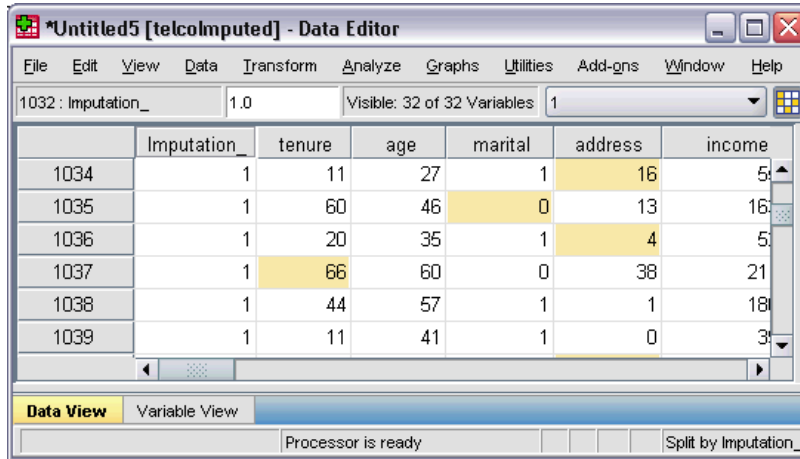
Figure 3-7  
*Data Editor with imputation markings OFF*



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

View  
Mark Imputed Data...

Figure 3-8  
Data Editor with imputation markings ON

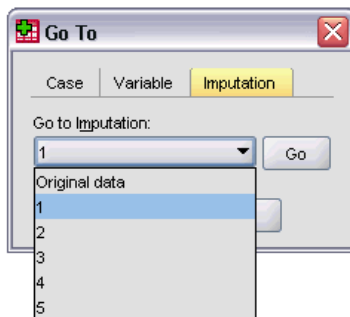


Alternatively, you can turn on markings by clicking the imputation marking button at the right edge of the edit bar in Data View of the Data Editor.

### ***Moving Between Imputations***

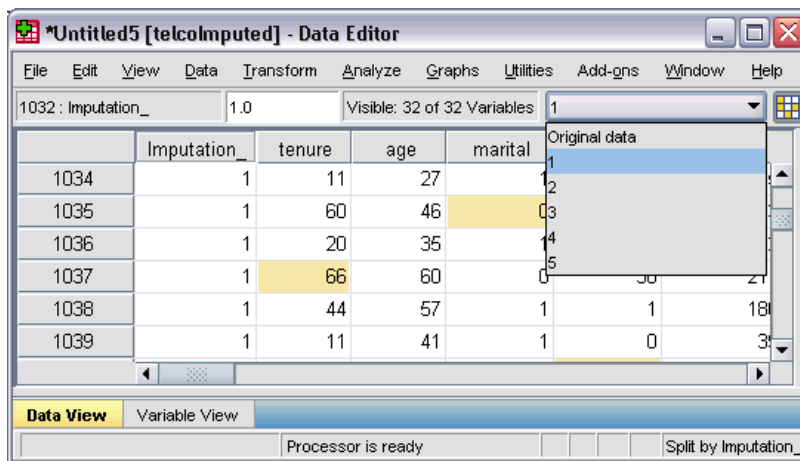
- ▶ From the menus choose:  
Edit  
Go to Imputation...
- ▶ Select the imputation (or Original data) from the drop-down list.

Figure 3-9  
Go To dialog box



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

Figure 3-10  
Data Editor with imputation markings ON



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***

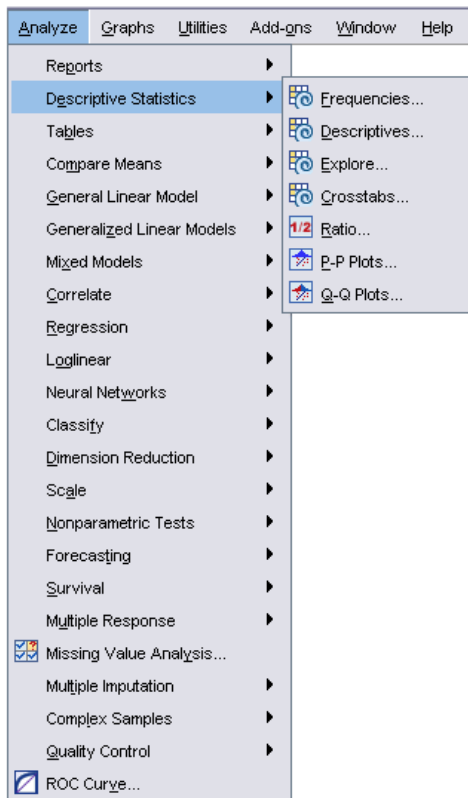
Sometimes you will need to perform transformations on imputed data. For example, you may want to take the log of all values of a salary variable and save the result in a new variable. A value computed using imputed data will be treated as imputed if it differs from the value computed using the original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

Figure 3-11  
Analyze menu with imputation markings ON



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 output.

- **Pooling of Tabular Output.** By default, when you run a supported procedure on a multiple imputation (MI) dataset, 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 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 datasets, at the levels of pooling specified for each piece of output.

#### **Frequencies**

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

**Crosstabs**

- The Crosstabulation table supports Count at Naïve pooling.

**Means**

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

**Independent-Samples T Test**

- 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 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 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, GLM Multivariate, and GLM Repeated**

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

**Partial Correlations**

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

#### **Binomial Test**

- 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 Descriptives table supports Means and N at Naïve pooling.



**One-Sample Kolmogorov-Smirnov Test**

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

**Two-Independent-Samples Tests**

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

**Two-Related-Samples Tests**

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

**Tests for Several Related Samples**

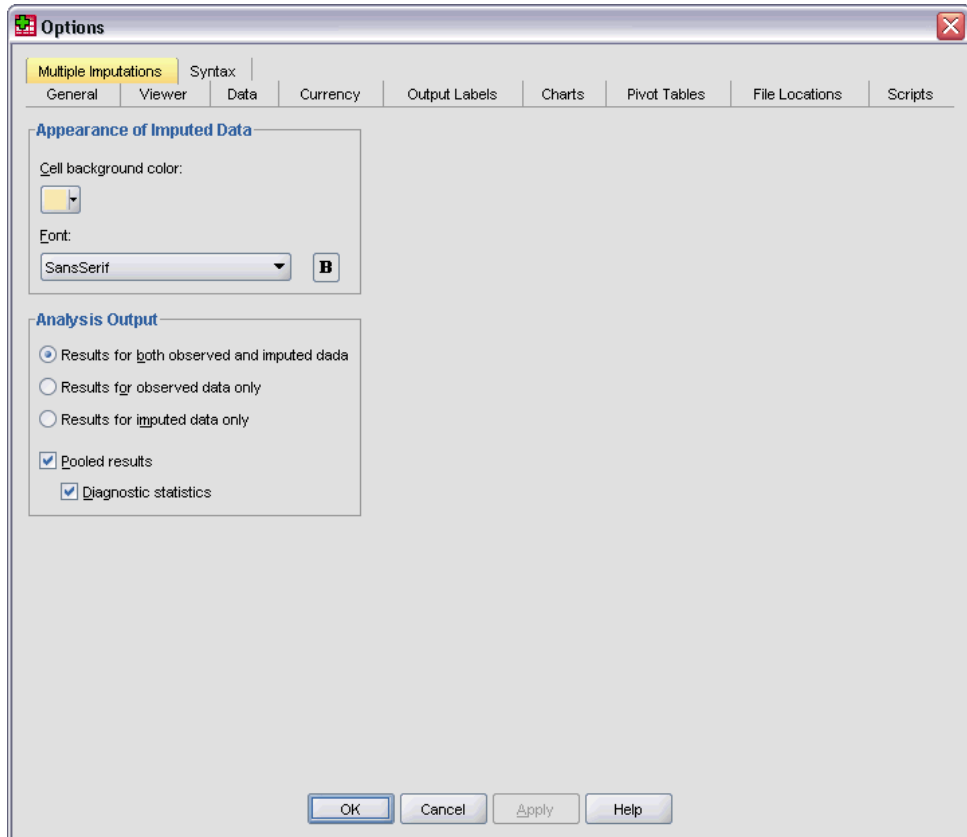
- The Ranks table supports Mean Rank at Naïve pooling.

**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.

## Multiple Imputation Options

Figure 3-12  
Options dialog box: Multiple Imputations tab



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

**Appearance 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.

# ***Part II: Examples***

# ***Missing Value Analysis***

## ***Describing the Pattern of Missing Data***

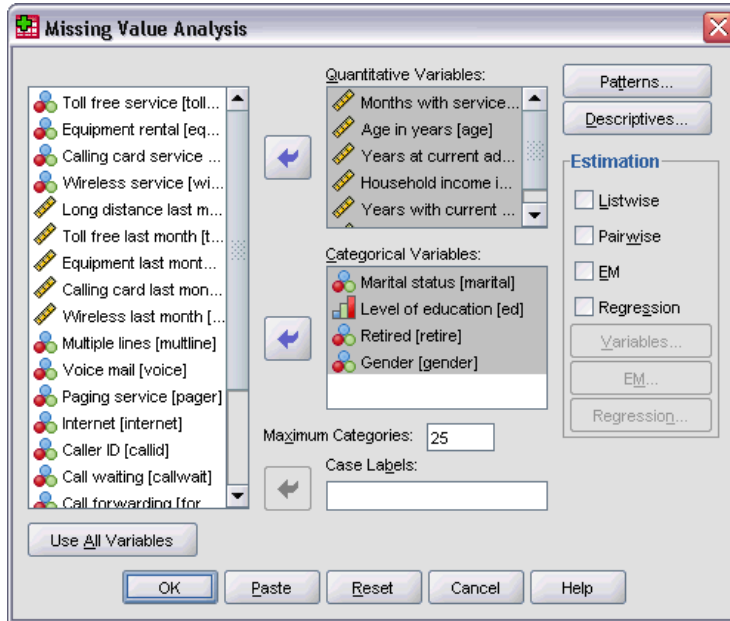
A telecommunications provider wants to better understand service usage patterns in its customer database. The company wants to ensure that the data are missing completely at random before running further analyses.

A random sample from the customer database is contained in *telco\_missing.sav*. For more information, see [Sample Files](#) in Appendix A on p. 100.

## ***Running the Analysis to Display Descriptive Statistics***

- ▶ To run the Missing Value Analysis, from the menus choose:
  - Analyze
  - Missing Value Analysis...

Figure 4-1  
Missing Value Analysis dialog box

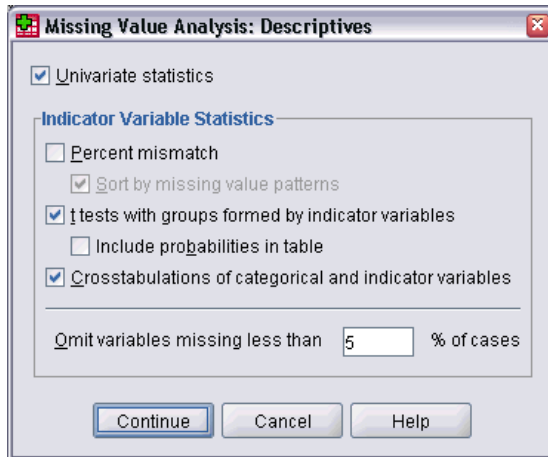


- ▶ Select *Marital status [marital]*, *Level of education [ed]*, *Retired [retire]*, and *Gender [gender]* as the categorical variables.
- ▶ Select *Months with service [tenure]* through *Number of people in household [reside]* as quantitative (scale) variables.

At this point, you could run the procedure and obtain univariate statistics, but we are going to select additional descriptive statistics.

- ▶ Click **Descriptives**.

Figure 4-2  
Missing Value Analysis: Descriptives dialog box



In the Descriptives dialog box, you can specify various descriptive statistics to display in the output. The default univariate statistics can help you to determine the general extent of the missing data, but the indicator variable statistics offer more information about how the pattern of missing data in one variable may affect the values of another variable.

- ▶ Select t tests with groups formed by indicator variables.
- ▶ Select Crosstabulations of categorical and indicator variables.
- ▶ Click Continue.
- ▶ In the main Missing Value Analysis dialog box, click OK.

## ***Evaluating the Descriptive Statistics***

For this example, the output includes:

- Univariate statistics

- Table of separate-variance *t* tests, including subgroup means when another variable is present or missing
- Tables for each categorical variable showing frequencies of missing data for each category by each quantitative (scale) variable

Figure 4-3  
Univariate statistics table

|         | N   | Mean    | Std. Deviation | Missing |         | No. of Extremes <sup>a</sup> |      |
|---------|-----|---------|----------------|---------|---------|------------------------------|------|
|         |     |         |                | Count   | Percent | Low                          | High |
| tenure  | 968 | 35.56   | 21.268         | 32      | 3.2     | 0                            | 0    |
| age     | 975 | 41.75   | 12.573         | 25      | 2.5     | 0                            | 0    |
| address | 850 | 11.47   | 9.965          | 150     | 15.0    | 0                            | 9    |
| income  | 821 | 71.1462 | 83.14424       | 179     | 17.9    | 0                            | 71   |
| employ  | 904 | 11.00   | 10.113         | 96      | 9.6     | 0                            | 15   |
| reside  | 966 | 2.32    | 1.431          | 34      | 3.4     | 0                            | 33   |
| marital | 885 |         |                | 115     | 11.5    |                              |      |
| ed      | 965 |         |                | 35      | 3.5     |                              |      |
| retire  | 916 |         |                | 84      | 8.4     |                              |      |
| gender  | 958 |         |                | 42      | 4.2     |                              |      |

a. Number of cases outside the range (Q1 - 1.5\*IQR, Q3 + 1.5\*IQR).

The univariate statistics provide your first look, variable by variable, at the extent of missing data. The number of nonmissing values for each variable appears in the *N* column, and the number of missing values appears in the *Missing Count* column. The *Missing Percent* column displays the percentage of cases with missing values and provides a good measure for comparing the extent of missing data among variables. *income* (*Household income in thousands*) has the greatest number of cases with missing values (17.9%), while *age* (*Age in years*) has the least (2.5%). *income* also has the greatest number of extreme values.



Figure 4-4  
Separate-variance *t* tests table

|               | tenure        | age   | address | income  | employ  | reside |
|---------------|---------------|-------|---------|---------|---------|--------|
| address       | t             | .4    | .3      | .3      | 1.4     | 1.0    |
|               | df            | 202.2 | 192.5   | .       | 313.6   | 191.1  |
|               | # Present     | 819   | 832     | 850     | 693     | 766    |
|               | # Missing     | 149   | 143     | 0       | 128     | 138    |
|               | Mean(Present) | 35.68 | 41.79   | 11.47   | 74.0779 | 11.20  |
| Mean(Missing) | 34.91         | 41.49 | .       | 55.2734 | 9.86    |        |
| income        | t             | -5.0  | -8.3    | -3.9    | .       | -5.9   |
|               | df            | 249.5 | 222.8   | 191.1   | .       | 203.3  |
|               | # Present     | 793   | 801     | 693     | 821     | 741    |
|               | # Missing     | 175   | 174     | 157     | 0       | 163    |
|               | Mean(Present) | 33.93 | 40.01   | 10.67   | 71.1462 | 9.91   |
| Mean(Missing) | 42.97         | 49.73 | 14.97   | .       | 15.93   |        |
| employ        | t             | -1.0  | -.4     | -.7     | .5      | -.3    |
|               | df            | 110.5 | 110.2   | 97.6    | 114.9   | .      |
|               | # Present     | 877   | 881     | 766     | 741     | 904    |
|               | # Missing     | 91    | 94      | 84      | 80      | 0      |
|               | Mean(Present) | 35.34 | 41.69   | 11.37   | 71.4953 | 11.00  |
| Mean(Missing) | 37.70         | 42.27 | 12.32   | 67.9125 | .       |        |
| marital       | t             | .0    | 1.8     | 1.2     | -.8     | .9     |
|               | df            | 148.1 | 149.5   | 138.8   | 121.2   | 128.3  |
|               | # Present     | 856   | 862     | 748     | 728     | 805    |
|               | # Missing     | 112   | 113     | 102     | 93      | 99     |
|               | Mean(Present) | 35.56 | 42.00   | 11.61   | 70.3887 | 11.10  |
| Mean(Missing) | 35.57         | 39.85 | 10.43   | 77.0753 | 10.17   |        |
| retire        | t             | -.6   | -.4     | -.4     | .3      | .2     |
|               | df            | 95.4  | 94.4    | 84.0    | 93.2    | .      |
|               | # Present     | 888   | 893     | 777     | 751     | 904    |
|               | # Missing     | 80    | 82      | 73      | 70      | 0      |
|               | Mean(Present) | 35.44 | 41.70   | 11.42   | 71.3356 | 11.00  |
| Mean(Missing) | 36.89         | 42.29 | 11.96   | 69.1143 | .       |        |

The separate-variance *t* tests table can help to identify variables whose pattern of missing values may be influencing the quantitative (scale) variables. The *t* test is computed using an indicator variable that specifies whether a variable is present or missing for an individual case. The subgroup means for the indicator variable are also tabulated. Note that an indicator variable is created only if a variable has missing values in at least 5% of the cases.

It appears that older respondents are less likely to report income levels. When *income* is missing, the mean *age* is 49.73, compared to 40.01 when *income* is nonmissing. In fact, the missingness of *income* seems to affect the means of several of the quantitative (scale) variables. This is one indication that the data may not be missing completely at random.

Figure 4-5  
Crosstabulation for Marital status [marital]

|         |         |          | Total | Unmarried | Married | Missing |
|---------|---------|----------|-------|-----------|---------|---------|
|         |         |          |       |           |         | SysMis  |
| address | Present | Count    | 850   | 390       | 358     | 102     |
|         |         | Percent  | 85.0  | 85.5      | 83.4    | 88.7    |
|         | Missing | % SysMis | 15.0  | 14.5      | 16.6    | 11.3    |
| income  | Present | Count    | 821   | 380       | 348     | 93      |
|         |         | Percent  | 82.1  | 83.3      | 81.1    | 80.9    |
|         | Missing | % SysMis | 17.9  | 16.7      | 18.9    | 19.1    |
| employ  | Present | Count    | 904   | 418       | 387     | 99      |
|         |         | Percent  | 90.4  | 91.7      | 90.2    | 86.1    |
|         | Missing | % SysMis | 9.6   | 8.3       | 9.8     | 13.9    |
| retire  | Present | Count    | 916   | 423       | 392     | 101     |
|         |         | Percent  | 91.6  | 92.8      | 91.4    | 87.8    |
|         | Missing | % SysMis | 8.4   | 7.2       | 8.6     | 12.2    |

The crosstabulations of categorical variables versus indicator variables show information similar to that found in the separate-variance  $t$  test table. Indicator variables are once again created, except this time they are used to calculate frequencies in every category for each categorical variable. The values can help you determine whether there are differences in missing values among categories.

Looking at the table for *marital* (*Marital status*), the number of missing values in the indicator variables do not appear to vary much between *marital* categories. Whether someone is married or unmarried does not seem to affect whether data are missing for any of the quantitative (scale) variables. For example, unmarried people reported *address* (*Years at current address*) 85.5% of the time, and married people reported the same variable 83.4% of the time. The difference is minimal and likely due to chance.

Figure 4-6  
Crosstabulation for Level of education [ed]

|         |         |          | Total | Did not complete high school | High school degree | Some college | College degree | Post-undergraduate degree | Missing |
|---------|---------|----------|-------|------------------------------|--------------------|--------------|----------------|---------------------------|---------|
|         |         |          |       |                              |                    |              |                |                           | SysMis  |
| address | Present | Count    | 850   | 163                          | 240                | 175          | 186            | 56                        | 30      |
|         |         | Percent  | 85.0  | 83.2                         | 85.7               | 88.4         | 81.9           | 87.5                      | 85.7    |
|         | Missing | % SysMis | 15.0  | 16.8                         | 14.3               | 11.6         | 18.1           | 12.5                      | 14.3    |
| income  | Present | Count    | 821   | 155                          | 229                | 165          | 193            | 50                        | 29      |
|         |         | Percent  | 82.1  | 79.1                         | 81.8               | 83.3         | 85.0           | 78.1                      | 82.9    |
|         | Missing | % SysMis | 17.9  | 20.9                         | 18.2               | 16.7         | 15.0           | 21.9                      | 17.1    |
| employ  | Present | Count    | 904   | 178                          | 254                | 178          | 204            | 60                        | 30      |
|         |         | Percent  | 90.4  | 90.8                         | 90.7               | 89.9         | 89.9           | 93.8                      | 85.7    |
|         | Missing | % SysMis | 9.6   | 9.2                          | 9.3                | 10.1         | 10.1           | 6.2                       | 14.3    |
| marital | Present | Count    | 885   | 193                          | 278                | 148          | 184            | 52                        | 30      |
|         |         | Percent  | 88.5  | 98.5                         | 99.3               | 74.7         | 81.1           | 81.2                      | 85.7    |
|         | Missing | % SysMis | 11.5  | 1.5                          | .7                 | 25.3         | 18.9           | 18.8                      | 14.3    |
| retire  | Present | Count    | 916   | 180                          | 259                | 180          | 207            | 60                        | 30      |
|         |         | Percent  | 91.6  | 91.8                         | 92.5               | 90.9         | 91.2           | 93.8                      | 85.7    |
|         | Missing | % SysMis | 8.4   | 8.2                          | 7.5                | 9.1          | 8.8            | 6.2                       | 14.3    |

Now consider the crosstabulation for *ed* (*Level of education*). If a respondent has at least some college education, a response for marital status is more likely to be missing. At least 98.5% of the respondents with no college education reported marital status. On the other hand, only 81.1% of those with a college degree reported marital status. The number is even lower for those with some college education but no degree.

**Figure 4-7**  
*Crosstabulation for Retired [retire]*

|         |         |          | Total | No   | Yes  | Missing<br>SysMis |
|---------|---------|----------|-------|------|------|-------------------|
| address | Present | Count    | 850   | 744  | 33   | 73                |
|         |         | Percent  | 85.0  | 85.0 | 80.5 | 86.9              |
|         | Missing | % SysMis | 15.0  | 15.0 | 19.5 | 13.1              |
| income  | Present | Count    | 821   | 732  | 19   | 70                |
|         |         | Percent  | 82.1  | 83.7 | 46.3 | 83.3              |
|         | Missing | % SysMis | 17.9  | 16.3 | 53.7 | 16.7              |
| employ  | Present | Count    | 904   | 864  | 40   | 0                 |
|         |         | Percent  | 90.4  | 98.7 | 97.6 | .0                |
|         | Missing | % SysMis | 9.6   | 1.3  | 2.4  | 100.0             |
| marital | Present | Count    | 885   | 777  | 38   | 70                |
|         |         | Percent  | 88.5  | 88.8 | 92.7 | 83.3              |
|         | Missing | % SysMis | 11.5  | 11.2 | 7.3  | 16.7              |

A more drastic difference can be seen in *retire (Retired)*. Those who are retired are much less likely to report their income compared to those who are not retired. Only 46.3% of the retired customers reported income level, while the percentage of those who are not retired and reported income level was 83.7.

**Figure 4-8**  
*Crosstabulation for Gender [gender]*

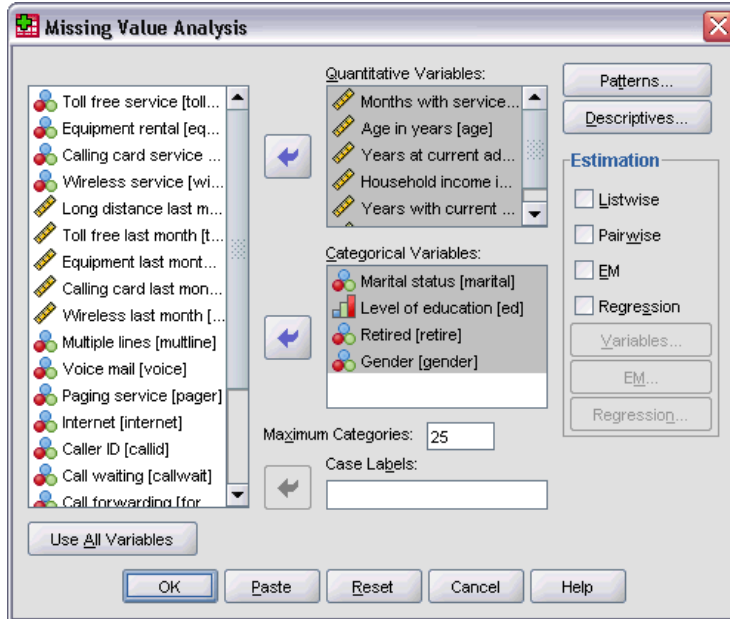
|         |         |          | Total | Male | Female | Missing |
|---------|---------|----------|-------|------|--------|---------|
|         |         |          |       |      |        | SysMis  |
| address | Present | Count    | 850   | 363  | 456    | 31      |
|         |         | Percent  | 85.0  | 78.6 | 91.9   | 73.8    |
|         | Missing | % SysMis | 15.0  | 21.4 | 8.1    | 26.2    |
| income  | Present | Count    | 821   | 381  | 406    | 34      |
|         |         | Percent  | 82.1  | 82.5 | 81.9   | 81.0    |
|         | Missing | % SysMis | 17.9  | 17.5 | 18.1   | 19.0    |
| employ  | Present | Count    | 904   | 412  | 457    | 35      |
|         |         | Percent  | 90.4  | 89.2 | 92.1   | 83.3    |
|         | Missing | % SysMis | 9.6   | 10.8 | 7.9    | 16.7    |
| marital | Present | Count    | 885   | 400  | 445    | 40      |
|         |         | Percent  | 88.5  | 86.6 | 89.7   | 95.2    |
|         | Missing | % SysMis | 11.5  | 13.4 | 10.3   | 4.8     |
| retire  | Present | Count    | 916   | 420  | 461    | 35      |
|         |         | Percent  | 91.6  | 90.9 | 92.9   | 83.3    |
|         | Missing | % SysMis | 8.4   | 9.1  | 7.1    | 16.7    |

Another discrepancy is apparent for *gender (Gender)*. Address information is missing more often for males than for females. Although these discrepancies could be due to chance, it seems unlikely. The data do not appear to be missing completely at random.

We will look at the patterns of missing data to explore this further.

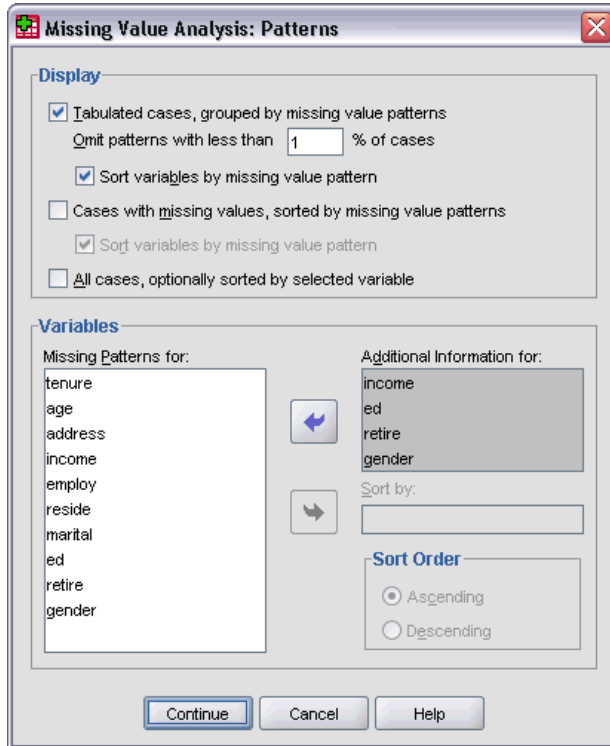
## Rerunning the Analysis to Display Patterns

Figure 4-9  
Missing Value Analysis dialog box



- ▶ Recall the Missing Value Analysis dialog box. The dialog remembers the variables used in the previous analysis. Do not change them.
- ▶ Click Patterns.

Figure 4-10  
Missing Value Analysis Patterns dialog box



In the Patterns dialog box, you can select various patterns tables. We are going to display tabulated patterns grouped by missing values patterns. Because the missing patterns in *ed* (*Level of education*), *retire* (*Retired*), and *gender* (*Gender*) seemed to influence the data, we will choose to display additional information for these variables. We will also include additional information for *income* (*Household income in thousands*) because of its large number of missing values.

- ▶ Select Tabulated cases, grouped by missing value patterns.
- ▶ Select *income*, *ed*, *retire*, and *gender* and add them to the Additional Information For list.
- ▶ Click Continue.
- ▶ In the main Missing Value Analysis dialog box, click OK.

## Evaluating the Patterns Table

Figure 4-11  
Tabulated patterns table

| Number of Cases | Missing Patterns <sup>a</sup> |        |        |    |        |        |        |         |         |        | Complete if ... <sup>b</sup> | income <sup>c</sup> | ed <sup>d</sup>              |                    |              |                |                           | retire <sup>d</sup> |     | gender <sup>d</sup> |        |
|-----------------|-------------------------------|--------|--------|----|--------|--------|--------|---------|---------|--------|------------------------------|---------------------|------------------------------|--------------------|--------------|----------------|---------------------------|---------------------|-----|---------------------|--------|
|                 | age                           | reside | tenure | ed | gender | retire | employ | marital | address | income |                              |                     | Did not complete high school | High school degree | Some college | College degree | Post-undergraduate degree | No                  | Yes | Male                | Female |
| 475             |                               |        |        |    |        |        |        |         |         |        | 475                          | 76.5853             | 99                           | 157                | 87           | 101            | 31                        | 463                 | 12  | 201                 | 274    |
| 109             |                               |        |        |    |        |        |        |         | X       | X      | 584                          | .                   | 27                           | 35                 | 19           | 17             | 11                        | 95                  | 14  | 47                  | 62     |
| 16              |                               |        |        |    |        |        |        | X       | X       |        | 687                          | .                   | 5                            | 9                  | 0            | 1              | 1                         | 12                  | 4   | 12                  | 4      |
| 87              |                               |        |        |    |        |        |        | X       |         |        | 562                          | 54.4368             | 21                           | 27                 | 9            | 24             | 6                         | 85                  | 2   | 66                  | 21     |
| 13              |                               | X      |        |    |        |        |        |         |         |        | 488                          | 56.0000             | 4                            | 3                  | 2            | 3              | 1                         | 13                  | 0   | 4                   | 9      |
| 60              |                               | X      |        |    |        |        |        | X       |         |        | 535                          | 77.2167             | 1                            | 2                  | 27           | 24             | 6                         | 59                  | 1   | 35                  | 25     |
| 16              |                               |        |        | X  |        |        |        |         |         |        | 491                          | 47.8125             | 0                            | 0                  | 0            | 0              | 0                         | 16                  | 0   | 6                   | 10     |
| 17              |                               |        | X      |    |        |        |        |         |         |        | 492                          | 76.2353             | 2                            | 7                  | 3            | 4              | 1                         | 17                  | 0   | 7                   | 10     |
| 18              |                               |        |        |    | X      |        |        |         |         |        | 493                          | 54.1111             | 3                            | 7                  | 4            | 4              | 0                         | 17                  | 1   | 0                   | 0      |
| 16              |                               |        |        |    |        |        | X      |         | X       |        | 660                          | .                   | 0                            | 0                  | 7            | 8              | 1                         | 14                  | 2   | 6                   | 10     |
| 37              |                               |        |        |    |        | X      | X      |         |         |        | 520                          | 59.4595             | 9                            | 14                 | 5            | 8              | 1                         | 0                   | 0   | 15                  | 22     |

Patterns with less than 1% cases (10 or fewer) are not displayed.

a. Variables are sorted on missing patterns.

b. Number of complete cases if variables missing in that pattern (marked with X) are not used.

c. Means at each unique pattern

d. Frequency distribution at each unique pattern

The tabulated patterns table shows whether the data tend to be missing for multiple variables in individual cases. That is, it can help you determine if your data are jointly missing.

There are three patterns of jointly missing data that occur in more than 1% of the cases. The variables *employ* (*Years with current employer*) and *retire* (*Retired*) are missing together more often than the other pairs. This is not surprising because *retire* and *employ* record similar information. If you don't know if a respondent is retired, you probably also don't know the respondent's years with current employer.

The mean *income* (*Household income in thousands*) seems to vary considerably depending on the missing value pattern. In particular, the mean *Income* is much higher for 6% (60 out of 1000) of the cases, when *marital* (*Marital status*) is missing. (It is also higher when *tenure* (*Months with service*) is missing, but this pattern accounts for

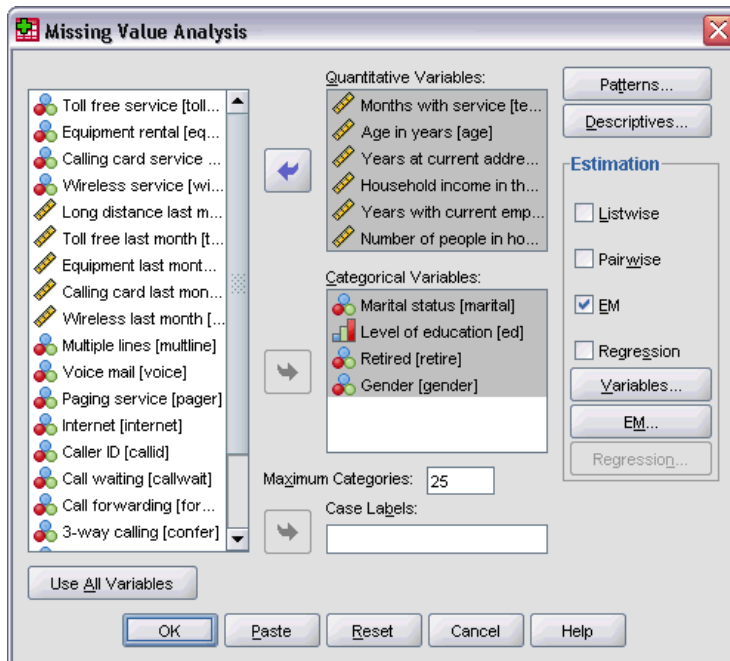


only 1.7% of the cases.) Remember that those with a higher level of education were less likely to respond to the question about marital status. You can see this trend in the frequencies shown for *ed* (*Level of education*). We might account for the increase in *income* by assuming that those with a higher level of education make more money and are less likely to report marital status.

Considering the descriptive statistics and patterns of missing data, we may be able to conclude that the data are not missing completely at random. We can confirm this conclusion through Little's MCAR test, which is printed with the EM estimates.

## Rerunning the Analysis for Little's MCAR Test

Figure 4-12  
Missing Value Analysis dialog box



- ▶ Recall the Missing Value Analysis dialog box.
- ▶ Click EM.

- ▶ Click OK.

Figure 4-13  
*EM means table*

| tenure | age   | address | income  | employ | reside |
|--------|-------|---------|---------|--------|--------|
| 36.12  | 41.91 | 11.58   | 77.3941 | 11.22  | 2.29   |

a. Little's MCAR test: Chi-Square = 179.836, DF = 107, Sig. = .000

The results of Little's MCAR test appear in footnotes to each EM estimate table. The null hypothesis for Little's MCAR test is that the data are missing completely at random (MCAR). Data are MCAR when the pattern of missing values does not depend on the data values. Because the significance value is less than 0.05 in our example, we can conclude that the data are *not* missing completely at random. This confirms the conclusion we drew from the descriptive statistics and tabulated patterns.

At this point, because the data are not missing completely at random, it is not safe to listwise delete cases with missing values or singly impute missing values. However, you can use [multiple imputation](#) to further analyze this dataset.

# ***Multiple Imputation***

## ***Using Multiple Imputation to Complete and Analyze a Dataset***

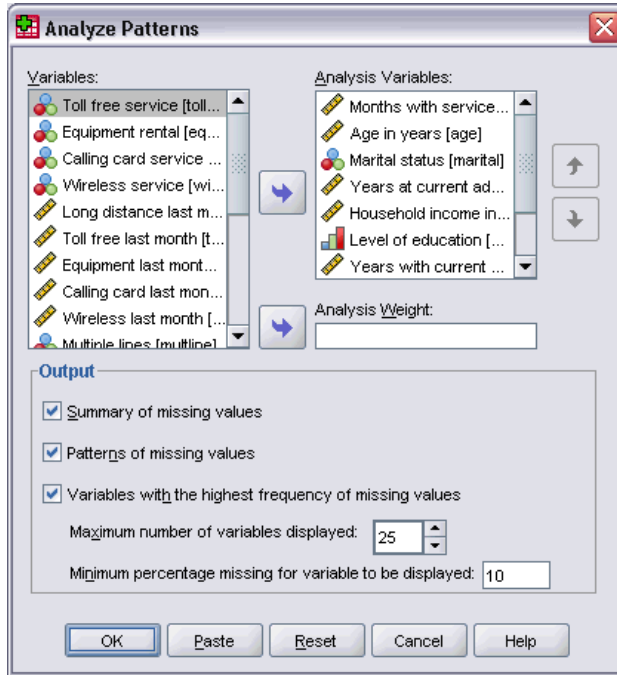
A telecommunications provider wants to better understand service usage patterns in its customer database. They have complete data for services used by their customers, but the demographic information collected by the company has a number of missing values. Moreover, these values are not missing completely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

A random sample from the customer database is contained in *telco\_missing.sav*. For more information, see [Sample Files](#) in Appendix A on p. 100.

### ***Analyze Patterns of Missing Values***

- ▶ As a first step, look at the patterns of missing data. From the menus choose:
  - Analyze
  - Multiple Imputation
  - Analyze Patterns...

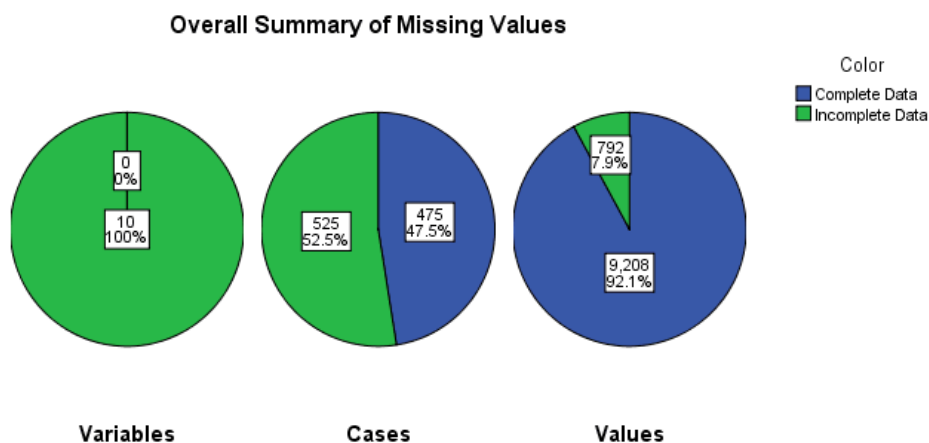
Figure 5-1  
*Analyze Patterns dialog*



- ▶ Select *Months with service [tenure]* through *Number of people in household [reside]* as analysis variables.

## Overall Summary

Figure 5-2  
Overall summary of missing values



The overall summary of missing values displays three pie charts that show different aspects of missing values in the data.

- The *Variables* chart shows that each of the 10 analysis variables has at least one missing value on a case.
- The *Cases* chart shows that 525 of the 1000 cases has at least one missing value on a variable.
- The *Values* chart shows that 792 of the 10,000 values (cases × variables) are missing.

Each case with missing values has, on average, missing values on roughly 1.5 of the 10 variables. This suggests that **listwise deletion** would lose much of the information in the dataset.

**Variable Summary**

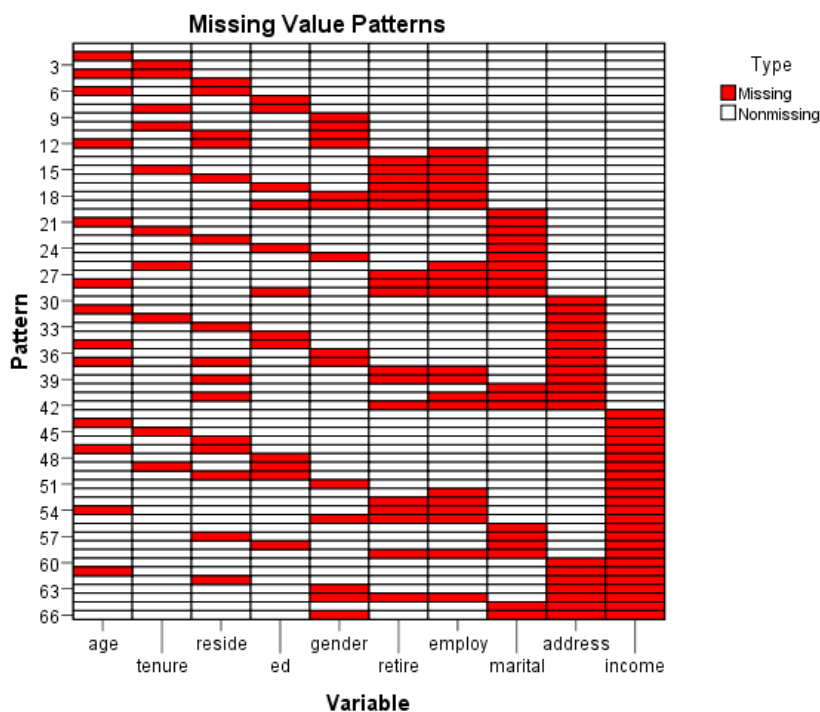
Figure 5-3  
Variable summary

|                               | Missing |         | Valid N | Mean    | Std.<br>Deviation |
|-------------------------------|---------|---------|---------|---------|-------------------|
|                               | N       | Percent |         |         |                   |
| Household income in thousands | 179     | 17.9%   | 821     | 71.1462 | 83.14424          |
| Years at current address      | 150     | 15.0%   | 850     | 11.47   | 9.965             |
| Marital status                | 115     | 11.5%   | 885     |         |                   |

The variable summary is displayed for variables with at least 10% missing values, and shows the number and percent of missing values for each variable in the table. It also displays the mean and standard deviation for the valid values of scale variables, and the number of valid values for all variables. *Household income in thousands*, *Years at current address*, and *Marital status* have the most missing values, in that order.

## Patterns

Figure 5-4  
Missing value patterns



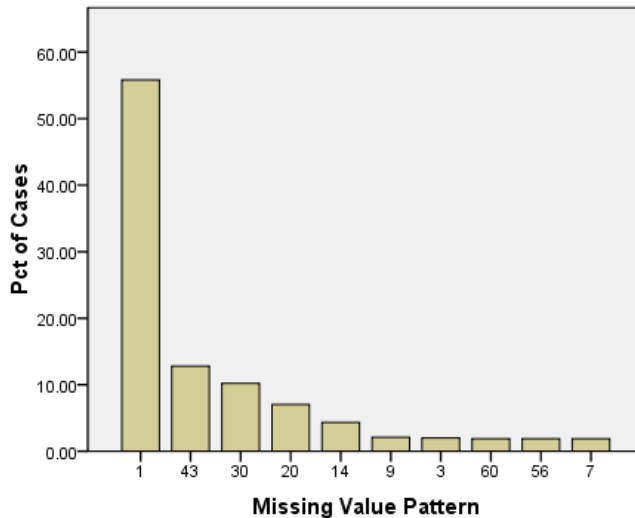
The patterns chart displays missing value patterns for the analysis variables. Each pattern corresponds to a group of cases with the same pattern of incomplete and complete data. For example, Pattern 1 represents cases which have no missing values, while Pattern 33 represents cases that have missing values on *reside* (*Number of people in household*) and *address* (*Years at current address*), and Pattern 66 represents cases which have missing values on *gender* (*Gender*), *marital* (*Marital status*), *address*, and *income* (*Household income in thousands*). A dataset can potentially have  $2^{\text{number of variables}}$  patterns. For 10 analysis variables this is  $2^{10}=1024$ ; however, only 66 patterns are represented in the 1000 cases in the dataset.

The chart orders analysis variables and patterns to reveal monotonicity where it exists. Specifically, variables are ordered from left to right in increasing order of missing values. Patterns are then sorted first by the last variable (nonmissing values

first, then missing values), then by the second to last variable, and so on, working from right to left. This reveals whether the monotone imputation method can be used for your data, or, if not, how closely your data approximate a monotone pattern. If the data are monotone, then all missing cells and nonmissing cells in the chart will be contiguous; that is, there will be no “islands” of nonmissing cells in the lower right portion of the chart and no “islands” of missing cells in the upper left portion of the chart.

This dataset is nonmonotone and there are many values that would need to be imputed in order to achieve monotonicity.

Figure 5-5  
Pattern frequencies



When patterns are requested a companion bar chart displays the percentage of cases for each pattern. This shows that over half of the cases in the dataset have Pattern 1, and the missing value patterns chart shows that this is the pattern for cases with no missing values. Pattern 43 represents cases with a missing value on *income*, Pattern 30 represents cases with a missing value on *address*, and Pattern 20 represents cases with a missing value on *marital*. The great majority of cases, roughly 4 in 5, are represented by these four patterns. Patterns 14, 60, and 56 are the only patterns among the ten most frequently occurring patterns to represent cases with missing values on more than one variable.



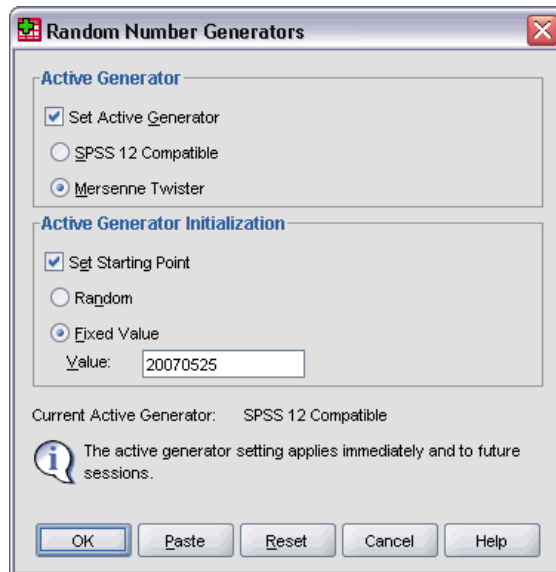
The analysis of missing patterns has not revealed any particular obstacles to multiple imputation, except that use of the monotone method will not really be feasible.

## Automatic Imputation of Missing Values

Now you are ready to begin imputing values; we'll start with a run with automatic settings, but before requesting imputations, we'll set the random seed. Setting the random seed allows you to replicate the analysis exactly.

- ▶ To set the random seed, from the menus choose:  
Transform  
Random Number Generators...

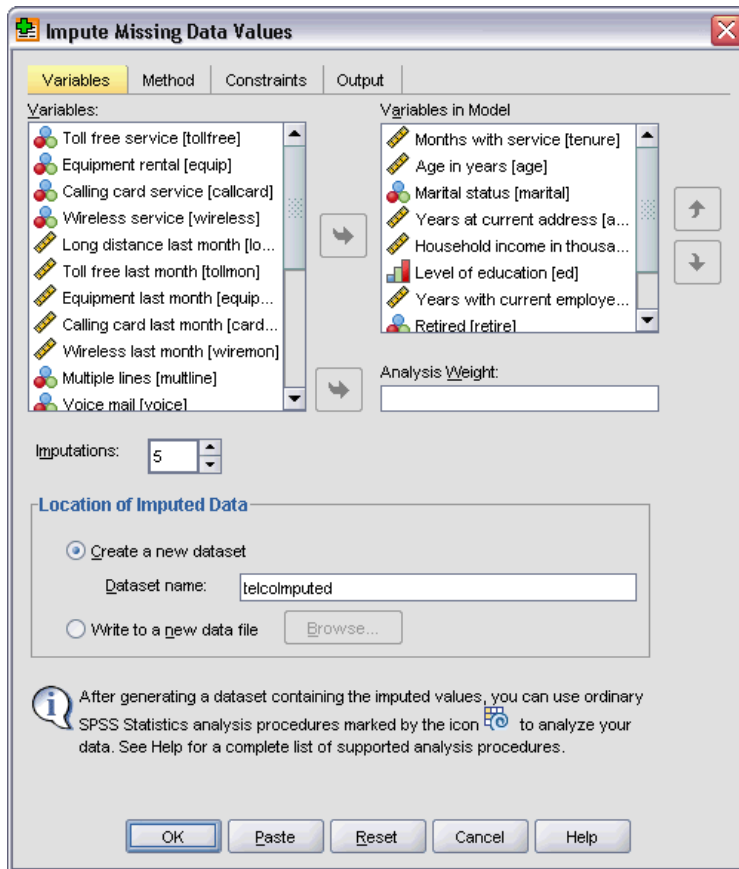
Figure 5-6  
*Random Number Generators dialog box*



- ▶ Select Set Active Generator.
- ▶ Select Mersenne Twister.
- ▶ Select Set Starting Point.

- ▶ Select Fixed Value, and type 20070525 as the value.
- ▶ Click OK.
- ▶ To multiply impute missing data values, from the menus choose:
  - Analyze
  - Multiple Imputation
  - Impute Missing Data Values...

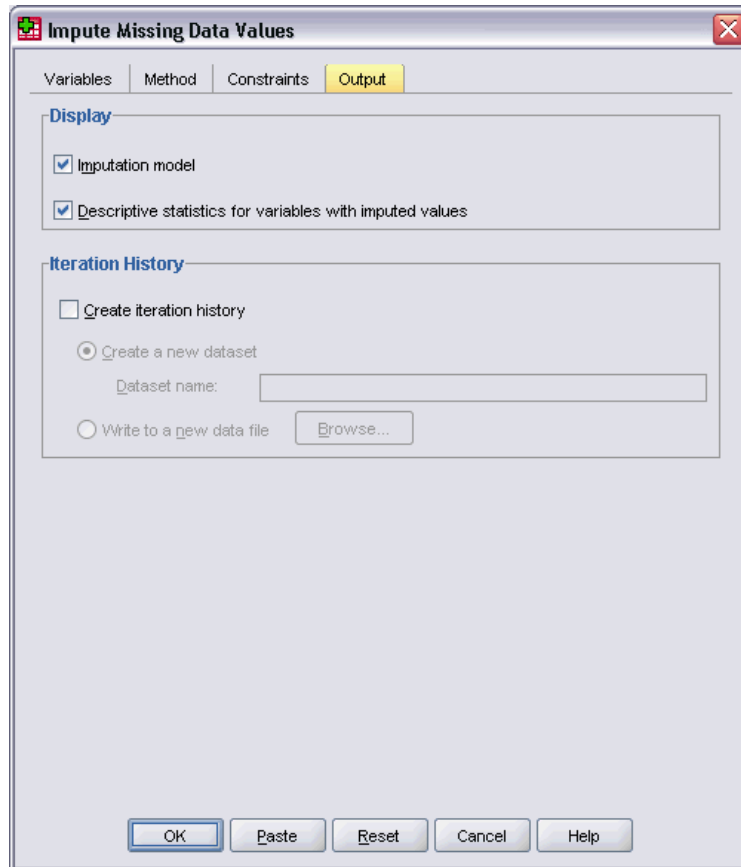
Figure 5-7  
Impute Missing Data Values dialog



- ▶ Select *Months with service [tenure]* through *Number of people in household [reside]* as variables in the imputation model.

- ▶ Type telcolimputed as the dataset to which imputed data should be saved.
- ▶ Click the Output tab.

Figure 5-8  
Output tab



- ▶ Select Descriptive statistics for variables with imputed values.
- ▶ Click OK.

## Imputation Specifications

Figure 5-9  
Imputation specifications

|                                      |                   |        |
|--------------------------------------|-------------------|--------|
| Imputation Method                    | Automatic         |        |
| Number of Imputations                |                   | 5      |
| Model for Scale Variables            | Linear Regression |        |
| Interactions Included in Models      | (none)            |        |
| Maximum Percentage of Missing Values |                   | 100.0% |

The imputation specifications table is a useful review of what you requested so that you can confirm that the specifications were correct.

## Imputation Results

Figure 5-10  
Imputation results

|   |                                      |  |
|---|--------------------------------------|--|
| Imputation Method                                 | Fully Conditional Specification      |  |
| Fully Conditional Specification Method Iterations |                                      | 10   |
| Dependent Variables                               | Imputed                              | tenure,age,marital,address,income,ed,employ,retire,gender,reside |
|   | Not Imputed(Too Many Missing Values) |  |
|   | Not Imputed(No Missing Values)       |  |
| Imputation Sequence                               |                                      | age,tenure,reside,ed,gender,retire,employ,marital,address,income |

The imputation results give an overview of what actually happened during the imputation process. Note in particular that:

- The imputation method in the specifications table was Automatic, and the method actually chosen by automatic method selection was Fully Conditional Specification.
- All requested variables were imputed.
- The imputation sequence is the order in which the variables appear on the *x*-axis on the Missing Value Patterns chart.

## Imputation Models

Figure 5-11  
Imputation models

|                               | Model               |   | Missing Values | Imputed Values |
|-------------------------------|---------------------|---|----------------|----------------|
|                               | Type                | Effects   |                |                |
| Age in years                  | Linear Regression   | ed,gender,retire,marital,tenure, reside,employ,address,income   | 25             | 125            |
| Months with service           | Linear Regression   | ed,gender,retire,marital,age, reside,employ,address,income      | 32             | 160            |
| Number of people in household | Linear Regression   | ed,gender,retire,marital,age, tenure,employ,address,income      | 34             | 170            |
| Level of education            | Logistic Regression | gender,retire,marital,age, tenure,reside,employ,address, income | 35             | 175            |
| Gender                        | Logistic Regression | ed,retire,marital,age,tenure, reside,employ,address,income      | 42             | 210            |
| Retired                       | Logistic Regression | ed,gender,marital,age,tenure, reside,employ,address,income      | 84             | 420            |
| Years with current employer   | Linear Regression   | ed,gender,retire,marital,age, tenure,reside,address,income      | 96             | 480            |
| Marital status                | Logistic Regression | ed,gender,retire,age,tenure, reside,employ,address,income       | 115            | 575            |
| Years at current address      | Linear Regression   | ed,gender,retire,marital,age, tenure,reside,employ,income       | 150            | 750            |
| Household income in thousands | Linear Regression   | ed,gender,retire,marital,age, tenure,reside,employ,address      | 179            | 895            |

The imputation models table gives further details about how each variable was imputed. Note in particular that:

- The variables are listed in the imputation sequence order.
- Scale variables are modeled with a linear regression, and categorical variables with a logistic regression.
- Each model uses all other variables as main effects.
- The number of missing values for each variable is reported, along with the total number of values imputed for that variable (number missing  $\times$  number of imputations).

### Descriptive Statistics

Figure 5-12  
Descriptive statistics for tenure (Months with service)

| Data                           | Imputation | N    | Mean  | Std. Deviation | Minimum | Maximum |
|--------------------------------|------------|------|-------|----------------|---------|---------|
| Original Data                  |            | 968  | 35.56 | 21.268         | 1.00    | 72.00   |
| Imputed Values                 | 1          | 32   | 36.06 | 24.218         | -6.72   | 90.02   |
|                                | 2          | 32   | 37.64 | 22.229         | -.19    | 88.03   |
|                                | 3          | 32   | 30.82 | 27.245         | -40.99  | 104.77  |
|                                | 4          | 32   | 39.97 | 20.585         | 1.29    | 80.50   |
|                                | 5          | 32   | 37.87 | 20.669         | 3.44    | 94.21   |
| Complete Data After Imputation | 1          | 1000 | 35.58 | 21.355         | -6.72   | 90.02   |
|                                | 2          | 1000 | 35.63 | 21.291         | -.19    | 88.03   |
|                                | 3          | 1000 | 35.41 | 21.484         | -40.99  | 104.77  |
|                                | 4          | 1000 | 35.70 | 21.251         | 1.00    | 80.50   |
|                                | 5          | 1000 | 35.63 | 21.243         | 1.00    | 94.21   |

Descriptive statistics tables show summaries for variables with imputed values. A separate table is produced for each variable. The types of statistics shown depend on whether the variable is scale or categorical.

Statistics for scale variables include the count, mean, standard deviation, minimum, and maximum, displayed for the original data, each set of imputed values, and each complete dataset (combining the original data and imputed values).

The descriptive statistics table for *tenure (Months with service)* shows means and standard deviations in each set of imputed values roughly equal to those in the original data; however, an immediate problem presents itself when you look at the minimum and see that negative values for *tenure* have been imputed.

**Figure 5-13**  
*Descriptive statistics for marital (Marital status)*

| Data                              | Imputation | Category | N   | Percent |
|-----------------------------------|------------|----------|-----|---------|
| Original Data                     |            | 0        | 456 | 51.5    |
|                                   |            | 1        | 429 | 48.5    |
| Imputed Values                    | 1          | 0        | 51  | 44.3    |
|                                   |            | 1        | 64  | 55.7    |
|                                   | 2          | 0        | 41  | 35.7    |
|                                   |            | 1        | 74  | 64.3    |
|                                   | 3          | 0        | 49  | 42.6    |
|                                   |            | 1        | 66  | 57.4    |
|                                   | 4          | 0        | 43  | 37.4    |
|                                   |            | 1        | 72  | 62.6    |
|                                   | 5          | 0        | 53  | 46.1    |
|                                   |            | 1        | 62  | 53.9    |
| Complete Data<br>After Imputation | 1          | 0        | 507 | 50.7    |
|                                   |            | 1        | 493 | 49.3    |
|                                   | 2          | 0        | 497 | 49.7    |
|                                   |            | 1        | 503 | 50.3    |
|                                   | 3          | 0        | 505 | 50.5    |
|                                   |            | 1        | 495 | 49.5    |
|                                   | 4          | 0        | 499 | 49.9    |
|                                   |            | 1        | 501 | 50.1    |
|                                   | 5          | 0        | 509 | 50.9    |
|                                   |            | 1        | 491 | 49.1    |

For categorical variables, statistics include count and percent by category for the original data, imputed values, and complete data. The table for *marital (Marital status)* has an interesting result in that, for the imputed values, a greater proportion of the cases are estimated as being married than in the original data. This could be due to random variation; alternatively the chance of being missing may be related to value of this variable.

**Figure 5-14**  
*Descriptive statistics for income (Household income in thousands)*

| Data                           | Imputation | N    | Mean     | Std. Deviation | Minimum   | Maximum  |
|--------------------------------|------------|------|----------|----------------|-----------|----------|
| Original Data                  |            | 821  | 71.1462  | 83.14424       | 9.0000    | 944.0000 |
| Imputed Values                 | 1          | 179  | 87.6574  | 91.13179       | -189.1959 | 373.2412 |
|                                | 2          | 179  | 101.6724 | 94.20599       | -122.0010 | 346.4294 |
|                                | 3          | 179  | 100.9445 | 95.00789       | -127.8572 | 342.5208 |
|                                | 4          | 179  | 107.0787 | 90.23638       | -113.0959 | 369.9674 |
|                                | 5          | 179  | 101.1043 | 90.40865       | -167.6978 | 314.2533 |
| Complete Data After Imputation | 1          | 1000 | 74.1017  | 84.81851       | -189.1959 | 944.0000 |
|                                | 2          | 1000 | 76.6104  | 85.98067       | -122.0010 | 944.0000 |
|                                | 3          | 1000 | 76.4801  | 86.10024       | -127.8572 | 944.0000 |
|                                | 4          | 1000 | 77.5781  | 85.52821       | -113.0959 | 944.0000 |
|                                | 5          | 1000 | 76.5087  | 85.22154       | -167.6978 | 944.0000 |

Like *tenure*, and all the other scale variables, *income (Household income in thousands)* shows negative imputed values — clearly, we will need to run a custom model with constraints on certain variables. However, *income* shows other potential problems. The mean values for each imputation are considerably higher than for the original data, and the maximum values for each imputation are considerably lower than for the original data. The distribution of income tends to be highly right-skew, so this could be the source of the problem.

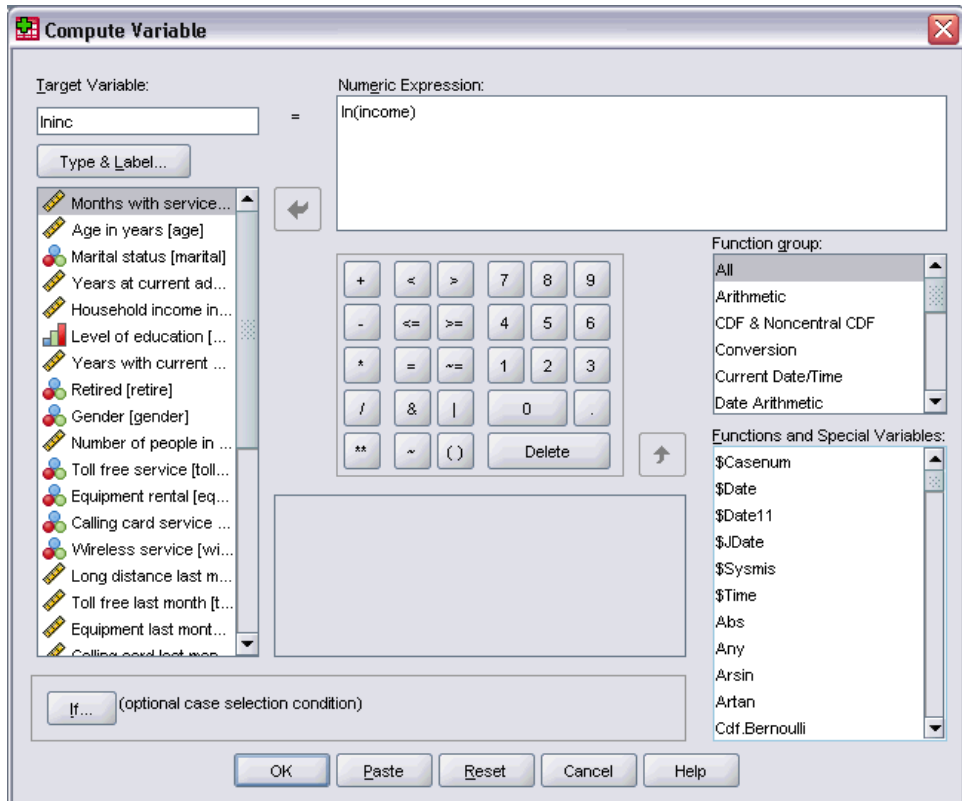
### ***Custom Imputation Model***

In order to prevent imputed values from falling outside the reasonable range of values for each variable, we'll specify a custom imputation model with constraints on the variables. Further, *Household income in thousands* is highly right-skew, and further analysis will likely use the logarithm of *income*, so it seems sensible to impute the log-income directly.

- ▶ Make sure the original dataset is active.
- ▶ To create a log-income variable, from the menus choose:  
 Transform  
 Compute Variable...

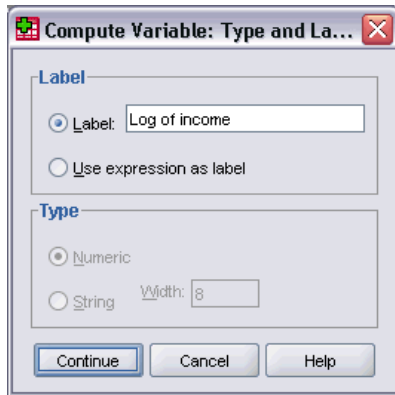


Figure 5-15  
Compute Variable dialog



- ▶ Type *lninc* as the target variable.
- ▶ Type *ln(income)* as the numeric expression.
- ▶ Click Type & Label..

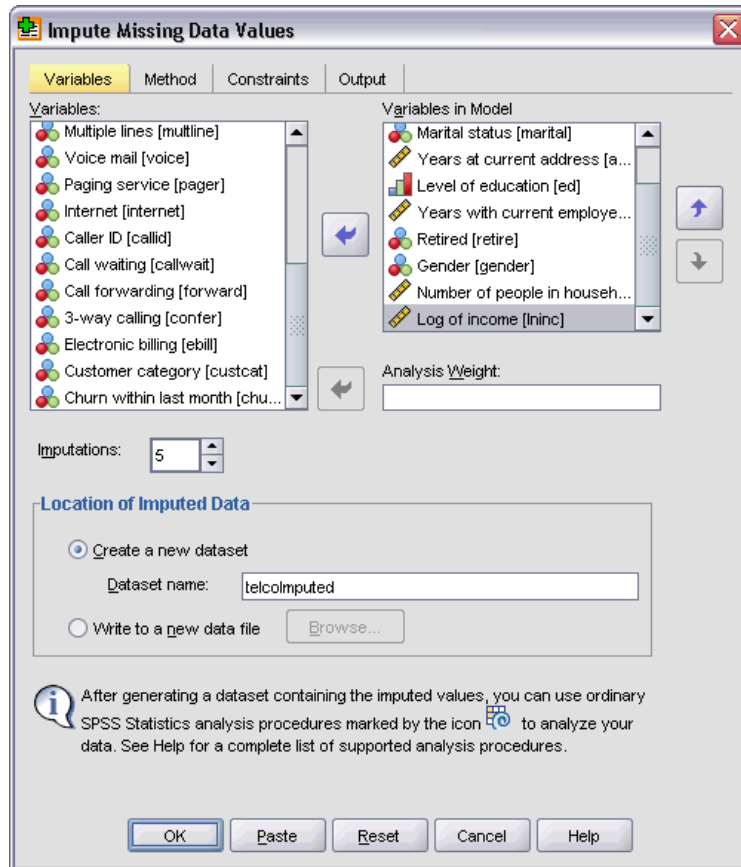
Figure 5-16  
*Type and Label dialog*



- ▶ Type *Log of income* as the label.
- ▶ Click Continue.
- ▶ Click OK in the Compute Variable dialog.

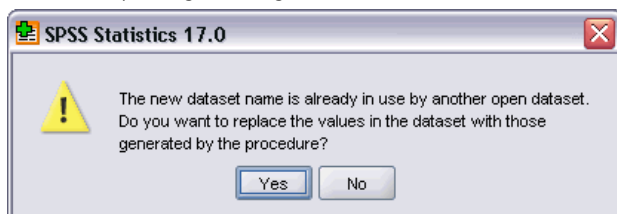
Figure 5-17

Variables tab with *Log of income* replacing *Household income in thousands* in the imputation model



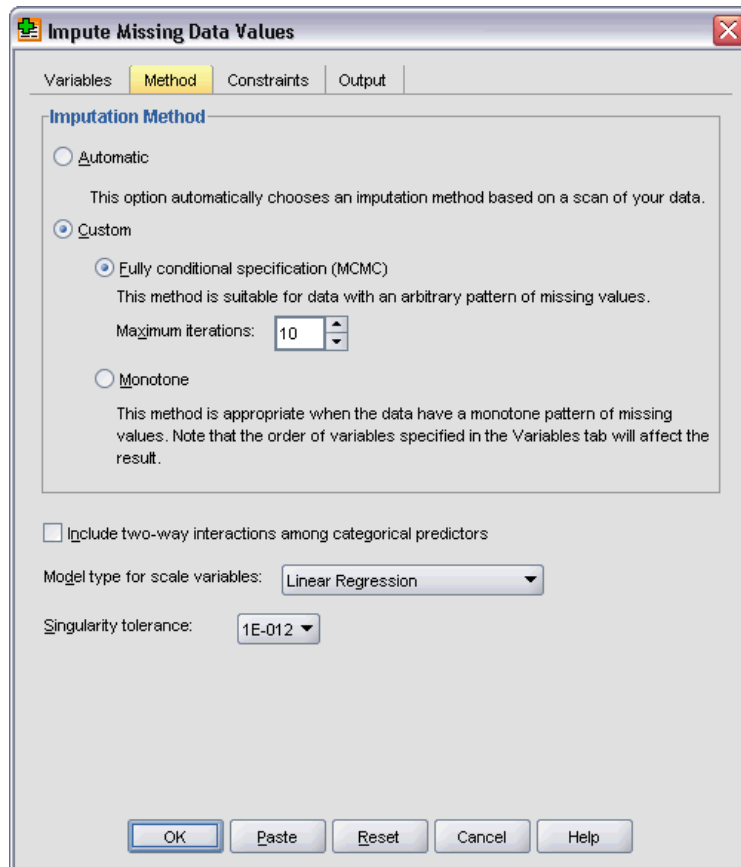
- ▶ Recall the Impute Missing Data Values dialog and click the Variables tab.
- ▶ Deselect *Household income in thousands [income]* and select *Log of income [lninc]* as variables in the model.
- ▶ Click the Method tab.

**Figure 5-18**  
*Alert for replacing existing dataset*



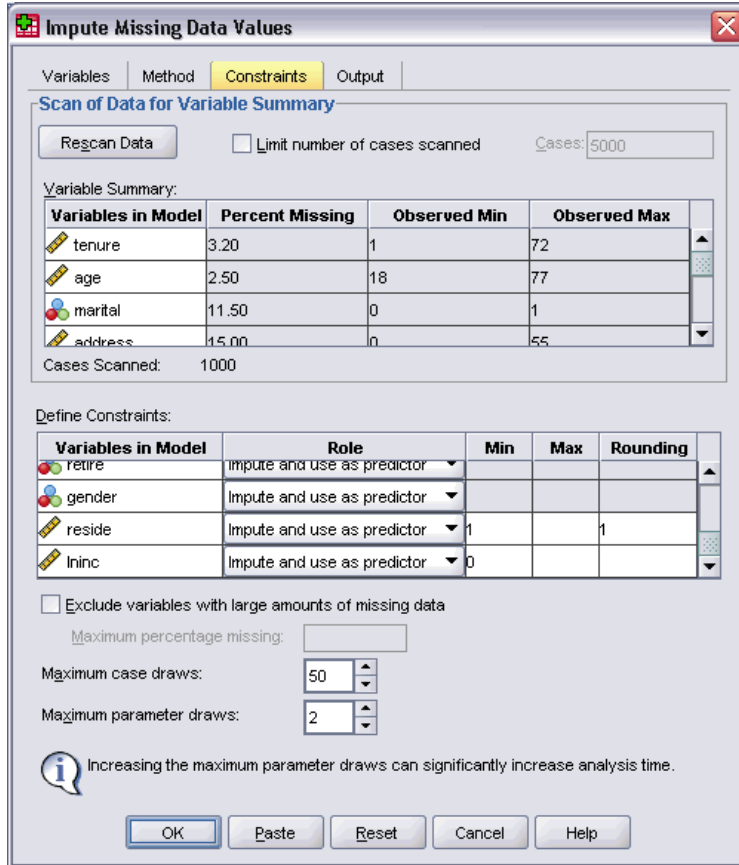
- ▶ Click Yes in the alert that appears.

Figure 5-19  
Method tab



- ▶ Select Custom and leave Fully conditional specification selected as the imputation method.
- ▶ Click the Constraints tab.

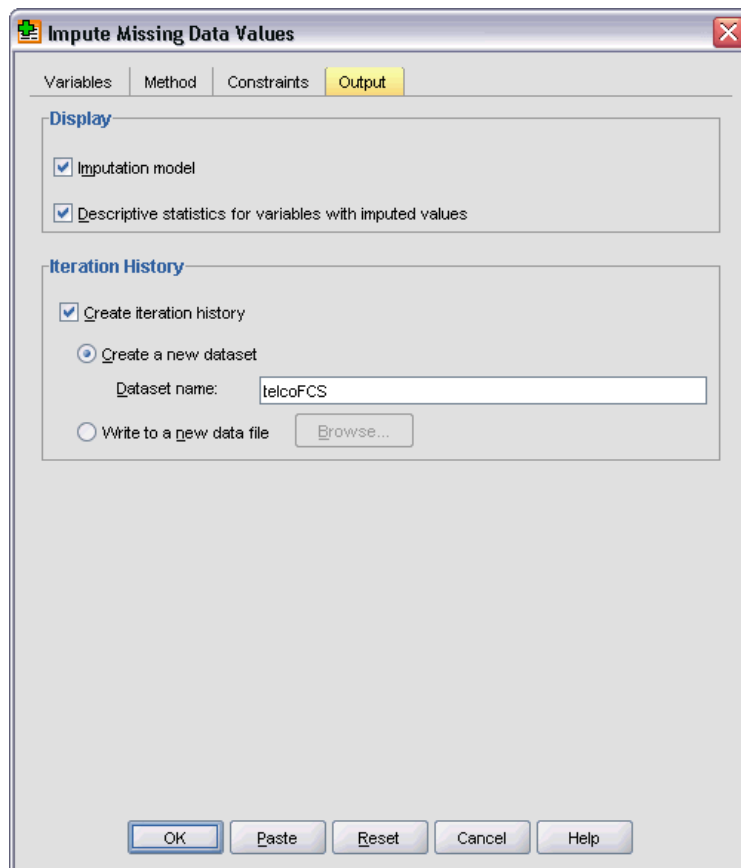
Figure 5-20  
Constraints tab



- ▶ Click Scan Data.
- ▶ In the Define Constraints grid, type 1 as the minimum value for *Months with service [tenure]*.
- ▶ Type 18 as the minimum value for *age (Age in years)*.
- ▶ Type 0 as the minimum value for *address (Years at current address)*.
- ▶ Type 0 as the minimum value for *employ (Years with current employer)*.

- ▶ Type 1 as the minimum value and 1 as the level of rounding for *reside* (*Number of people in household*). Note that while many of the other scale variables are reported in integer values, it is sensible to posit that someone has lived for 13.8 years at their current address, but not really to think that 2.2 people live there.
- ▶ Type 0 as the minimum value for *lninc* (*Log of income*).
- ▶ Click the Output tab.

Figure 5-21  
Output tab



- ▶ Select Create iteration history and type telcoFCS as the name of the new dataset.
- ▶ Click OK.

### Imputation Constraints

Figure 5-22  
Imputation constraints

|                               | Role in Imputation |           | Imputed Values |         |          |
|-------------------------------|--------------------|-----------|----------------|---------|----------|
|                               | Dependent          | Predictor | Minimum        | Maximum | Rounding |
| Months with service           | Yes                | Yes       | 1              | (none)  |          |
| Age in years                  | Yes                | Yes       | 18             | (none)  |          |
| Marital status                | Yes                | Yes       |                |         |          |
| Years at current address      | Yes                | Yes       | 0              | (none)  |          |
| Level of education            | Yes                | Yes       |                |         |          |
| Years with current employer   | Yes                | Yes       | 0              | (none)  |          |
| Retired                       | Yes                | Yes       |                |         |          |
| Gender                        | Yes                | Yes       |                |         |          |
| Number of people in household | Yes                | Yes       | 1              | (none)  | Integer  |
| Log of Income                 | Yes                | Yes       | 0              | (none)  |          |

The custom imputation model results in a new table that reviews the constraints placed upon the imputation model. Everything appears to be in accordance with your specifications.

### Descriptive Statistics

Figure 5-23  
Descriptive statistics for tenure (Months with service)

| Data                           | Imputation | N    | Mean  | Std. Deviation | Minimum | Maximum |
|--------------------------------|------------|------|-------|----------------|---------|---------|
| Original Data                  |            | 968  | 35.56 | 21.268         | 1.00    | 72.00   |
| Imputed Values                 | 1          | 32   | 37.90 | 17.621         | 6.40    | 86.94   |
|                                | 2          | 32   | 40.97 | 24.517         | 5.33    | 90.48   |
|                                | 3          | 32   | 37.57 | 19.913         | 9.97    | 93.52   |
|                                | 4          | 32   | 39.69 | 21.644         | 9.25    | 83.61   |
|                                | 5          | 32   | 39.85 | 20.093         | 7.21    | 81.37   |
| Complete Data After Imputation | 1          | 1000 | 35.64 | 21.158         | 1.00    | 86.94   |
|                                | 2          | 1000 | 35.73 | 21.387         | 1.00    | 90.48   |
|                                | 3          | 1000 | 35.63 | 21.220         | 1.00    | 93.52   |
|                                | 4          | 1000 | 35.69 | 21.282         | 1.00    | 83.61   |
|                                | 5          | 1000 | 35.70 | 21.235         | 1.00    | 81.37   |

The descriptive statistics table for *tenure (Months with service)* under the custom imputation model with constraints shows that the problem of negative imputed values for *tenure* has been solved.



Figure 5-24  
Descriptive statistics for marital (*Marital status*)

| Data           | Imputation                     | Category | N   | Percent |      |
|----------------|--------------------------------|----------|-----|---------|------|
| Original Data  |                                | 0        | 456 | 51.5    |      |
|                |                                | 1        | 429 | 48.5    |      |
| Imputed Values | 1                              | 0        | 46  | 40.0    |      |
|                |                                | 1        | 69  | 60.0    |      |
|                | 2                              | 0        | 43  | 37.4    |      |
|                |                                | 1        | 72  | 62.6    |      |
|                | 3                              | 0        | 60  | 52.2    |      |
|                |                                | 1        | 55  | 47.8    |      |
|                | 4                              | 0        | 45  | 39.1    |      |
|                |                                | 1        | 70  | 60.9    |      |
|                | 5                              | 0        | 49  | 42.6    |      |
|                |                                | 1        | 66  | 57.4    |      |
|                | Complete Data After Imputation | 1        | 0   | 502     | 50.2 |
|                |                                |          | 1   | 498     | 49.8 |
| 2              |                                | 0        | 499 | 49.9    |      |
|                |                                | 1        | 501 | 50.1    |      |
| 3              |                                | 0        | 516 | 51.6    |      |
|                |                                | 1        | 484 | 48.4    |      |
| 4              |                                | 0        | 501 | 50.1    |      |
|                |                                | 1        | 499 | 49.9    |      |
| 5              |                                | 0        | 505 | 50.5    |      |
|                |                                | 1        | 495 | 49.5    |      |

The table for *marital (Marital status)* now has an imputation (3) whose distribution is more in line with the original data, but the majority are still showing a greater proportion of the cases estimated as being married than in the original data. This could be due to random variation, but might require further study of the data to determine whether these values are not missing at random (MAR). We will not pursue this further here.

**Figure 5-25**  
*Descriptive statistics for lninc (Log of income)*

| Data                              | Imputation | N    | Mean   | Std. Deviation | Minimum | Maximum |
|-----------------------------------|------------|------|--------|----------------|---------|---------|
| Original Data                     |            | 821  | 3.9291 | .75305         | 2.1972  | 6.8501  |
| Imputed Values                    | 1          | 179  | 4.1816 | .94574         | 1.4428  | 6.6748  |
|                                   | 2          | 179  | 4.2562 | .98346         | 1.6633  | 6.8224  |
|                                   | 3          | 179  | 4.1743 | 1.01487        | 1.4443  | 6.8437  |
|                                   | 4          | 179  | 4.1774 | .82705         | 2.2532  | 6.2680  |
|                                   | 5          | 179  | 4.1894 | .96403         | 1.6667  | 6.6677  |
| Complete Data<br>After Imputation | 1          | 1000 | 3.9743 | .79638         | 1.4428  | 6.8501  |
|                                   | 2          | 1000 | 3.9876 | .80842         | 1.6633  | 6.8501  |
|                                   | 3          | 1000 | 3.9730 | .81107         | 1.4443  | 6.8501  |
|                                   | 4          | 1000 | 3.9735 | .77228         | 2.1972  | 6.8501  |
|                                   | 5          | 1000 | 3.9756 | .80064         | 1.6667  | 6.8501  |

Like *tenure*, and all the other scale variables, *lninc (Log of income)* does not show negative imputed values. Moreover, the mean values for imputations are closer to the mean for the original data than in the automatic imputation run — in the *income* scale, the mean for the original data for *lninc* is approximately  $e^{3.9291}=50.86$ , while the typical mean value among the imputations is very roughly  $e^{4.2}=66.69$ . Additionally, the maximum values for each imputation are closer to the maximum value for the original data.

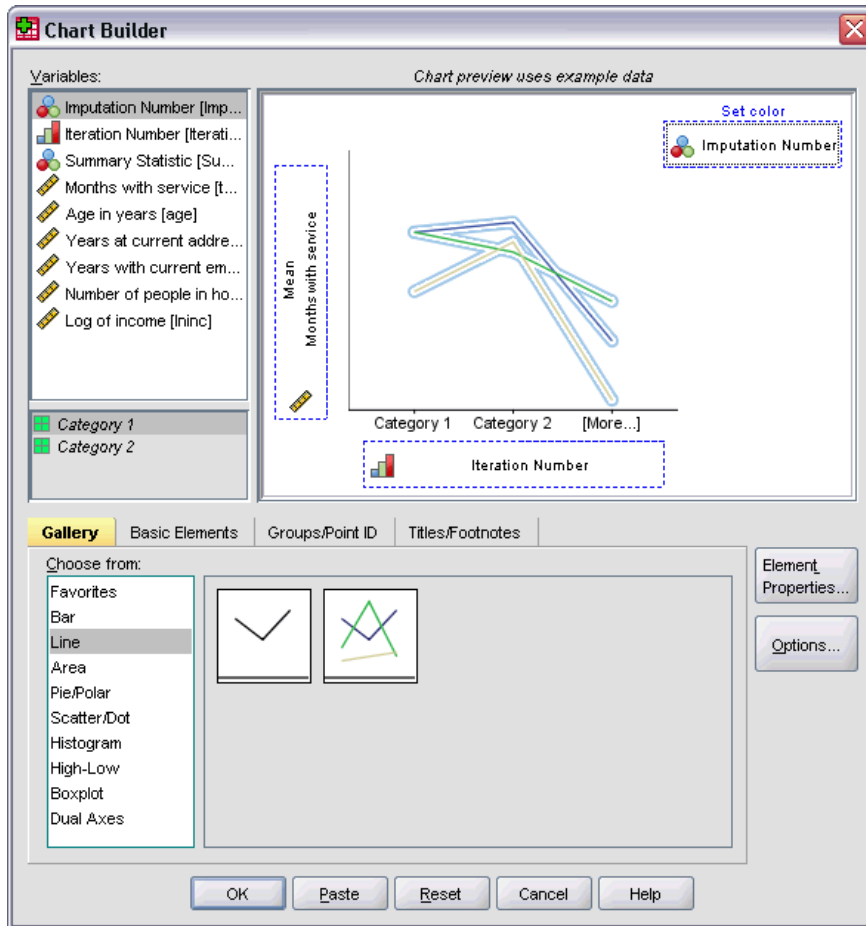
### **Checking FCS Convergence**

When using the fully conditional specification method, it's a good idea to check plots of the means and standard deviations by iteration and imputation for each scale dependent variable for which values are imputed in order to help assess model convergence.

- ▶ To create this type of chart, activate the *telcoFCS* dataset, and then from the menus choose:

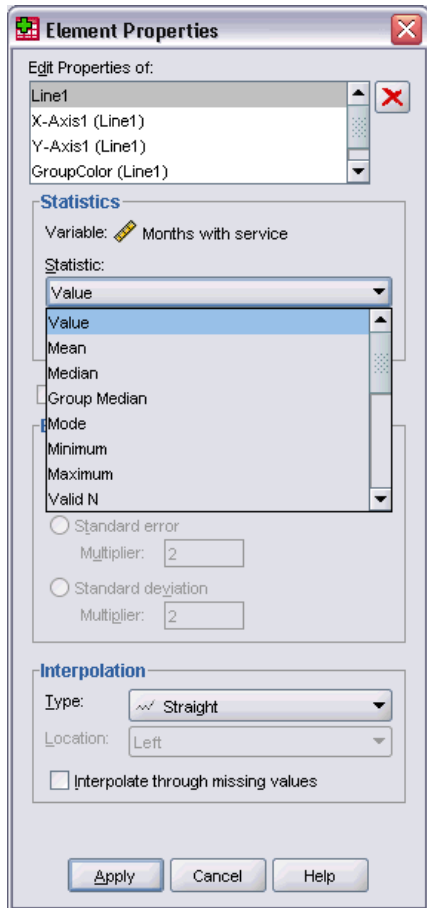
Graphs  
 Chart Builder...

Figure 5-26  
Chart Builder, Multiple Lines plot



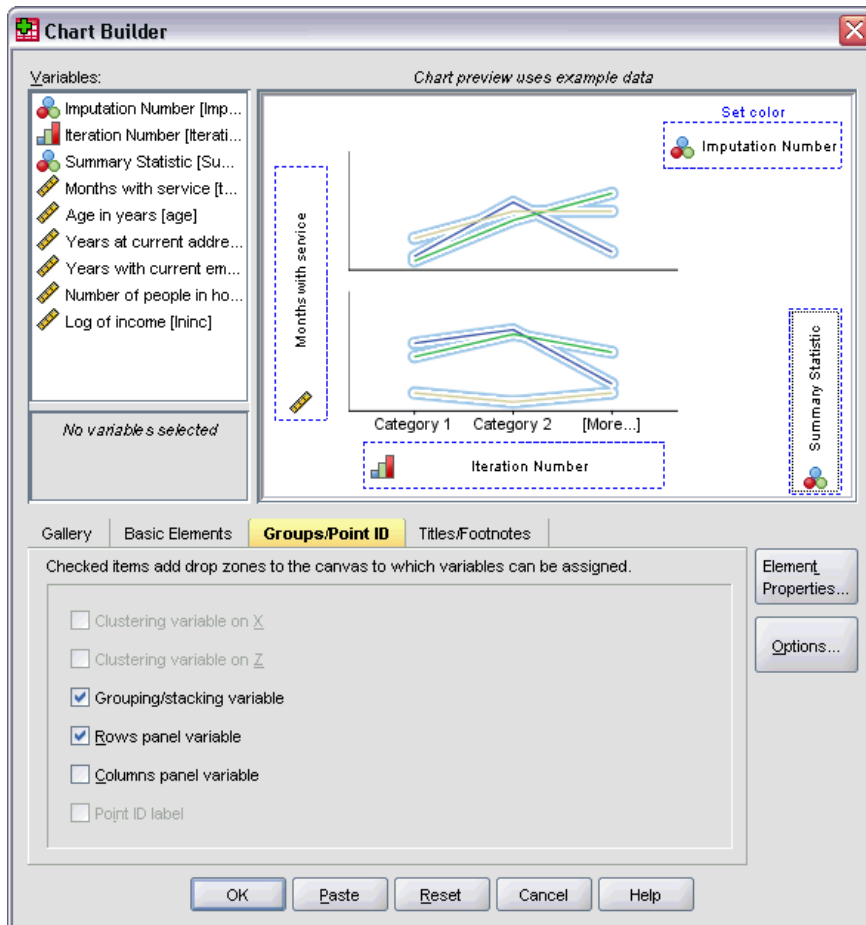
- ▶ Select the Line gallery and choose Multiple Line.
- ▶ Select *Months with service [tenure]* as the variable to plot on the Y axis.
- ▶ Select *Iteration Number [Iteration\_]* as the variable to plot on the X axis.
- ▶ Select *Imputation Number [Imputations\_]* as the variable to set colors by.

Figure 5-27  
Chart Builder, Element Properties



- ▶ In the Element Properties, select Value as the statistic to display.
- ▶ Click Apply.
- ▶ In the Chart Builder, click the Groups/Point ID tab.

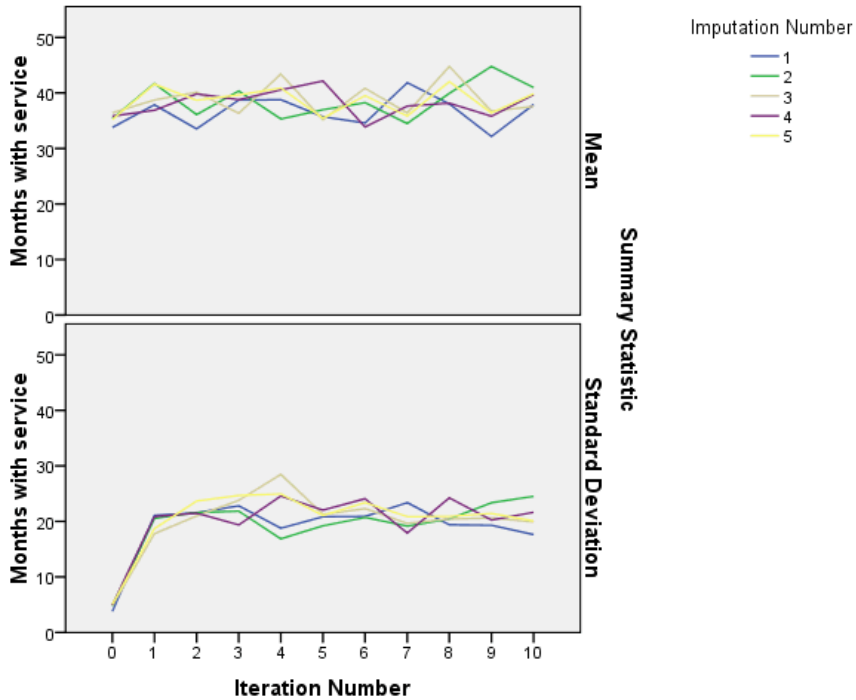
Figure 5-28  
Chart Builder, Groups/Point ID tab



- ▶ Select Rows panel variable.
- ▶ Select *Summary Statistic* [*SummaryStatistic\_*] as the panel variable.
- ▶ Click OK.

### FCS Convergence Charts

Figure 5-29  
FCS convergence chart



You have created a pair of multiple line charts, showing the mean and standard deviation of the imputed values of *Months with service [tenure]* at each iteration of the FCS imputation method for each of the 5 requested imputations. The purpose of this plot is to look for patterns in the lines. There should not be any, and these look suitably “random”. You can create similar plots for the other scale variables, and note that those plots also show no discernable patterns.

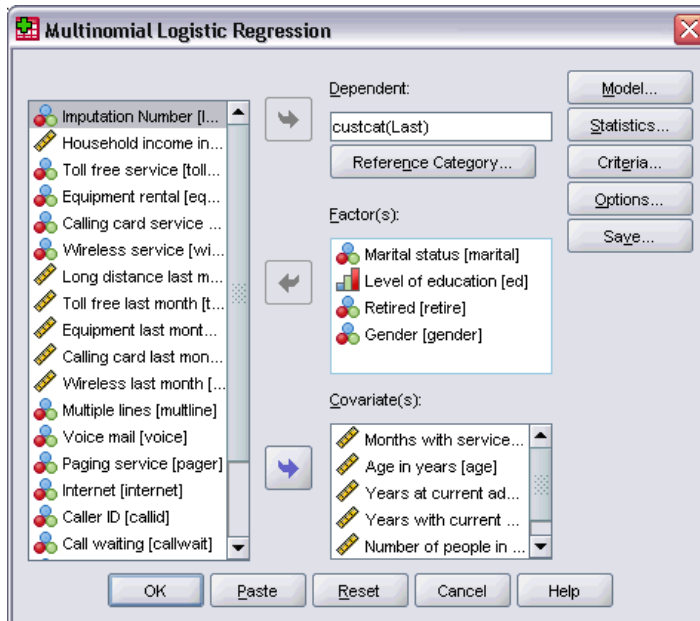
## Analyze Complete Data

Now that your imputed values appear to be satisfactory, you are ready to run an analysis on the “complete” data. The dataset contains a variable *Customer category* [*custcat*] that segments the customer base by service usage patterns, categorizing the customers into four groups. If you can fit a model using demographic information to predict group membership, you can customize offers for individual prospective customers.

- ▶ Activate the *telcoImputed* dataset. To create a multinomial logistic regression model for the complete data, from the menus choose:

```
Analyze
  Regression
    Multinomial Logistic...
```

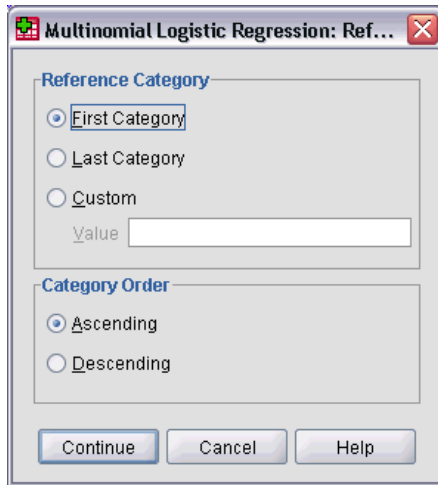
Figure 5-30  
*Multinomial Logistic Regression dialog*



- ▶ Select *Customer category* as the dependent variable.
- ▶ Select *Marital status*, *Level of education*, *Retired*, and *Gender* as factors.

- ▶ Select *Age in Years*, *Years at current address*, *Years with current employer*, *Number of people in household*, and *Log of income* as covariates.
- ▶ You want to compare other customers to those who subscribe to the Basic service, so select *Customer category* and click Reference category.

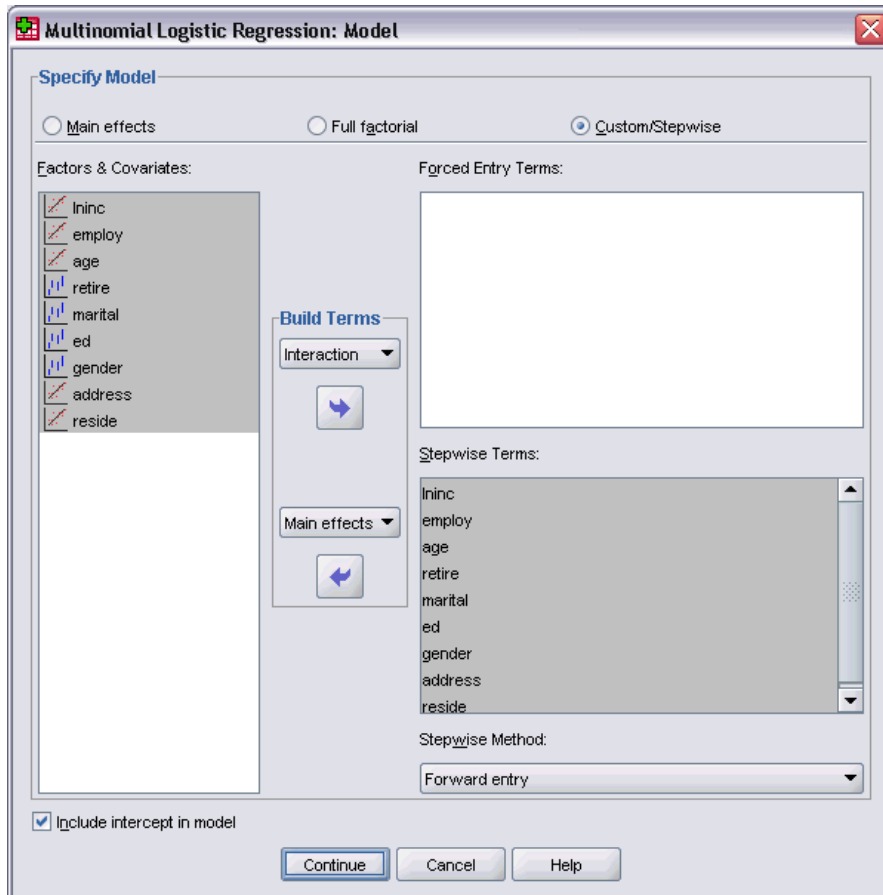
Figure 5-31  
Reference Category dialog box



- ▶ Select First category.
- ▶ Click Continue.
- ▶ Click Model in the Multinomial Logistic Regression dialog box.



Figure 5-32  
Model dialog box



- ▶ Select Custom/Stepwise.
- ▶ Select Main effects from the Stepwise Terms Build Term(s) dropdown.
- ▶ Select *lninc* through *reside* as Stepwise Terms.
- ▶ Click Continue.
- ▶ Click OK in the Multinomial Logistic Regression dialog box.

### Step Summary

Figure 5-33  
Step summary

| Imputation Number | Model | Action  | Effect(s) | Model Fitting Criteria | Effect Selection Tests  |    |      |
|-------------------|-------|---------|-----------|------------------------|-------------------------|----|------|
|                   |       |         |           | -2 Log Likelihood      | Chi-Square <sup>a</sup> | df | Sig. |
| Original data     | 0     | Entered | Intercept | 1353.555               | .                       |    |      |
|                   | 1     | Entered | ed        | 1260.972               | 92.583                  | 12 | .000 |
|                   | 2     | Entered | employ    | 1237.664               | 23.308                  | 3  | .000 |
|                   | 3     | Entered | marital   | 1229.808               | 7.856                   | 3  | .049 |
| 1                 | 0     | Entered | Intercept | 2762.531               | .                       |    |      |
|                   | 1     | Entered | ed        | 2608.189               | 154.342                 | 12 | .000 |
|                   | 2     | Entered | employ    | 2563.671               | 44.518                  | 3  | .000 |
|                   | 3     | Entered | reside    | 2549.200               | 14.470                  | 3  | .002 |
|                   | 4     | Entered | address   | 2541.050               | 8.151                   | 3  | .043 |
| 2                 | 0     | Entered | Intercept | 2762.531               | .                       |    |      |
|                   | 1     | Entered | ed        | 2603.940               | 158.591                 | 12 | .000 |
|                   | 2     | Entered | employ    | 2563.367               | 40.573                  | 3  | .000 |
|                   | 3     | Entered | marital   | 2545.743               | 17.624                  | 3  | .001 |
|                   | 4     | Entered | address   | 2536.532               | 9.211                   | 3  | .027 |
| 3                 | 0     | Entered | Intercept | 2762.531               | .                       |    |      |
|                   | 1     | Entered | ed        | 2600.074               | 162.457                 | 12 | .000 |
|                   | 2     | Entered | employ    | 2558.560               | 41.514                  | 3  | .000 |
|                   | 3     | Entered | marital   | 2546.062               | 12.499                  | 3  | .006 |
|                   | 4     | Entered | address   | 2536.348               | 9.714                   | 3  | .021 |
| 4                 | 0     | Entered | Intercept | 2762.531               | .                       |    |      |
|                   | 1     | Entered | ed        | 2601.616               | 160.915                 | 12 | .000 |
|                   | 2     | Entered | employ    | 2558.463               | 43.153                  | 3  | .000 |
|                   | 3     | Entered | marital   | 2543.747               | 14.716                  | 3  | .002 |
|                   | 4     | Entered | address   | 2533.341               | 10.406                  | 3  | .015 |
| 5                 | 0     | Entered | Intercept | 2762.531               | .                       |    |      |
|                   | 1     | Entered | ed        | 2604.773               | 157.759                 | 12 | .000 |
|                   | 2     | Entered | employ    | 2561.792               | 42.980                  | 3  | .000 |
|                   | 3     | Entered | marital   | 2549.096               | 12.696                  | 3  | .005 |

Stepwise Method: Forward Entry

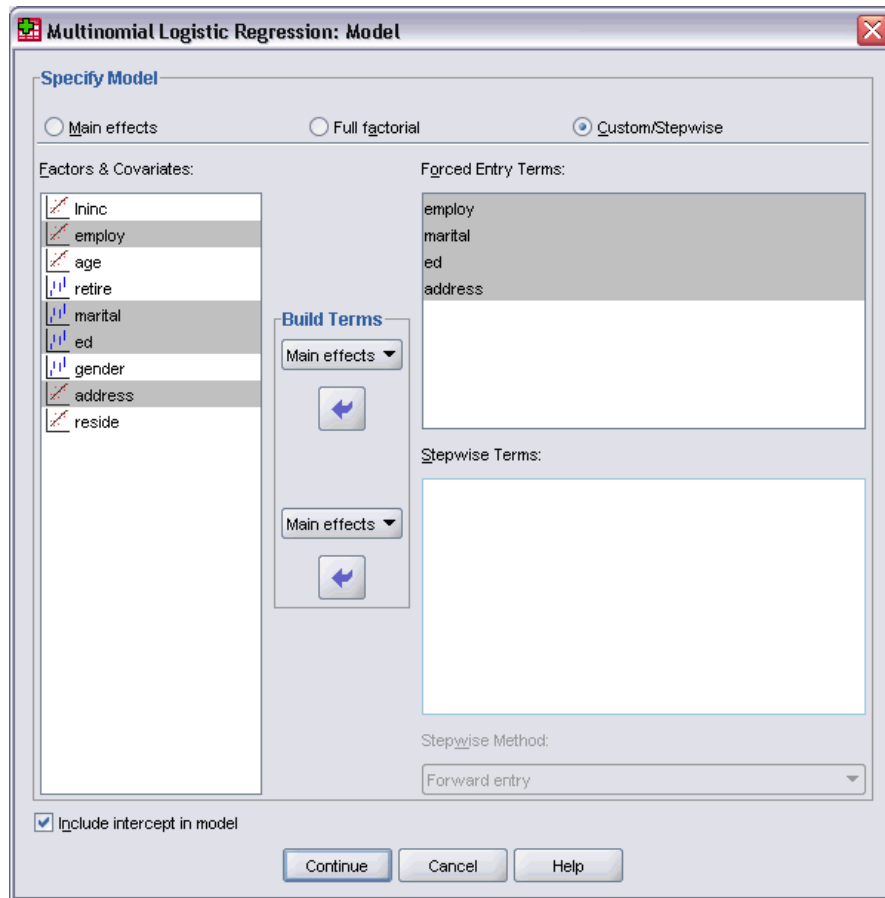
a. The chi-square for entry is based on the likelihood ratio test.

Multinomial Logistic Regression supports pooling of regression coefficients; however, you will note *all* tables in the output show the results for each imputation and the original data. This is because the file is split on *Imputation\_*, so all tables that honor the split variable will present the split file groups together in a single table.

You will also see that the Parameter Estimates table does not show pooled estimates; to answer why, look at the Step Summary. We requested stepwise selection of model effects, and the same set of effects was not chosen for all imputations, thus it is not possible to perform pooling. However, this still provides useful information because we see that *ed* (*Level of education*), *employ* (*Years with current employer*), *marital* (*Marital status*), and *address* (*Years at current address*) are frequently chosen by stepwise selection among the imputations. We will fit another model using just these predictors.

### Running the Model with a Subset of Predictors

Figure 5-34  
Model dialog



- ▶ Recall the Multinomial Logistic Regression dialog and click Model.
- ▶ Deselect the variables from the Stepwise Terms list.
- ▶ Select Main effects from the Forced Entry Terms Build Term(s) dropdown.
- ▶ Select *employ*, *marital*, *ed*, and *address* as Forced Entry Terms.
- ▶ Click Continue.

- Click OK in the Multinomial Logistic Regression dialog box.

### ***Pooled Parameter Estimates***

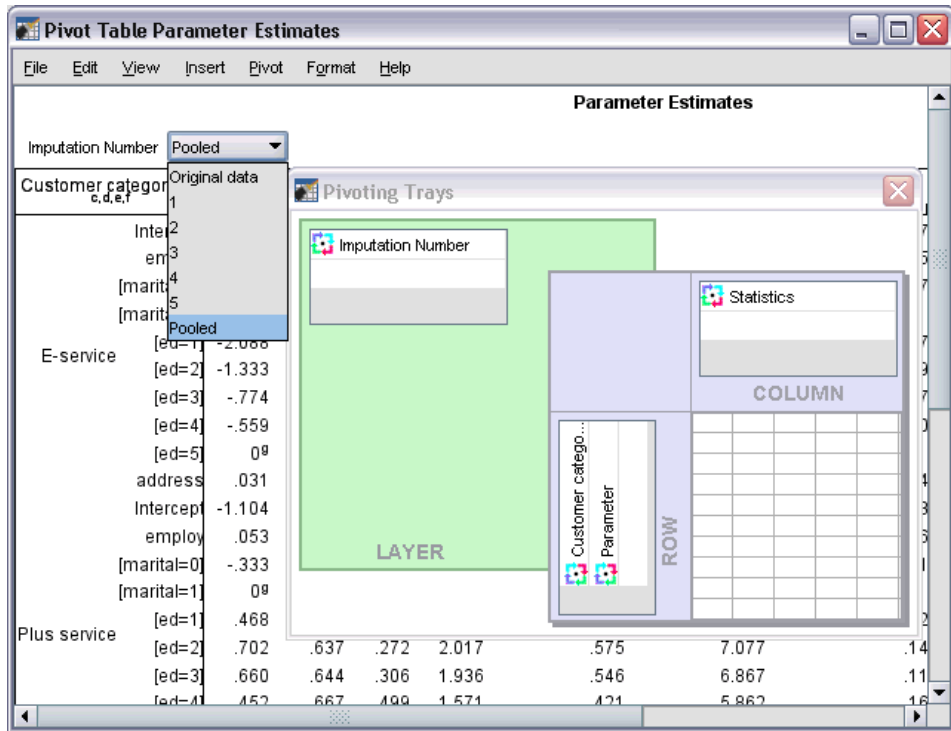
This table is rather large, but pivoting will give us a couple of different useful views of the output.

Figure 5-35  
*Pooled parameter estimates*

| Imputation Number | Customer category | Wald     | df   | Sig.  | Exp(B)  | Lower Bound |       |
|-------------------|-------------------|----------|------|-------|---------|-------------|-------|
| Original data     | E-service         | 1.137    | 1    | .286  |         |             |       |
|                   | Intercept         | 8.981    | 1    | .003  | 1.056   |             |       |
|                   | Internet          | 7.137    | 1    | .008  | .468    |             |       |
|                   | Mobile            |          | 0    |       |         |             |       |
|                   | Landline          | 11.503   | 1    | .001  | .103    |             |       |
|                   | Landline          | 7.256    | 1    | .007  | .191    |             |       |
|                   | Landline          | 1.707    | 1    | .191  | .428    |             |       |
|                   | Landline          | .794     | 1    | .373  | .562    |             |       |
|                   | Landline          |          | 0    |       |         |             |       |
|                   | Landline          | 2.651    | 1    | .104  | 1.029   |             |       |
| Plus service      | Intercept         | 2238.103 | 1    | .000  |         |             |       |
|                   | Internet          | 12.225   | 1    | .000  | 1.058   |             |       |
|                   | Mobile            | 4.680    | 1    | .031  | .578    |             |       |
|                   | Landline          |          | 0    |       |         |             |       |
|                   | Landline          | 1860.334 | 1    | .000  | 4.121E7 |             |       |
|                   | Landline          | 2037.269 | 1    | .000  | 3.441E7 |             |       |
|                   | Landline          | 1559.017 | 1    | .000  | 5.176E7 |             |       |
| Total service     | address           | .023     | .016 | 2.219 | 1       | .136        | 1.024 |
|                   | Intercept         | 1.266    | .556 | 5.177 | 1       | .023        |       |

- Activate (double-click) the table, then select Pivoting Trays from the context menu.

Figure 5-36  
Pooled parameter estimates



- ▶ Move *Imputation Number* from the Row into the Layer.
- ▶ Select Pooled from the Imputation Number dropdown list.

Figure 5-37  
Pooled parameter estimates

| Customer category <sup>a</sup> ,<br>b,c,d,e,f | Parameter   | Statistics     |            |      |        |                                    |             |                        |                            |                     |
|---|-------------|----------------|------------|------|--------|------------------------------------|-------------|------------------------|----------------------------|---------------------|
|   |             | B              | Std. Error | Sig. | Exp(B) | 95% Confidence Interval for Exp(B) |             | Fraction Missing Info. | Relative Increase Variance | Relative Efficiency |
|   |             |                |            |      |        | Lower Bound                        | Upper Bound |                        |                            |                     |
| E-service                                     | Intercept   | .553           | .435       | .204 |        |                                    |             | .076                   | .080                       | .985                |
|   | employ      | .030           | .012       | .014 | 1.030  | 1.006                              | 1.054       | .051                   | .052                       | .990                |
|   | [marital=0] | -.565          | .198       | .004 | .568   | .385                               | .839        | .076                   | .080                       | .985                |
|   | [marital=1] | 0 <sup>g</sup> |            |      |        |                                    |             |                        |                            |                     |
|   | [ed=1]      | -2.088         | .479       | .000 | .124   | .048                               | .317        | .079                   | .082                       | .985                |
|   | [ed=2]      | -1.333         | .454       | .004 | .264   | .108                               | .644        | .092                   | .098                       | .982                |
|   | [ed=3]      | -.774          | .458       | .092 | .461   | .187                               | 1.134       | .075                   | .079                       | .985                |
|   | [ed=4]      | -.559          | .466       | .231 | .572   | .229                               | 1.428       | .109                   | .116                       | .979                |
|   | [ed=5]      | 0 <sup>g</sup> |            |      |        |                                    |             |                        |                            |                     |
|   | address     | .031           | .012       | .009 | 1.032  | 1.008                              | 1.056       | .140                   | .153                       | .973                |
| Plus service                                  | Intercept   | -1.104         | .632       | .082 |        |                                    |             | .139                   | .152                       | .973                |
|   | employ      | .053           | .011       | .000 | 1.054  | 1.032                              | 1.076       | .060                   | .062                       | .988                |
|   | [marital=0] | -.333          | .179       | .063 | .717   | .505                               | 1.018       | .011                   | .012                       | .998                |
|   | [marital=1] | 0 <sup>g</sup> |            |      |        |                                    |             |                        |                            |                     |
|   | [ed=1]      | .468           | .638       | .464 | 1.597  | .455                               | 5.609       | .126                   | .136                       | .975                |
|   | [ed=2]      | .702           | .637       | .272 | 2.017  | .575                               | 7.077       | .140                   | .153                       | .973                |
|   | [ed=3]      | .660           | .644       | .306 | 1.936  | .546                               | 6.867       | .118                   | .127                       | .977                |
|   | [ed=4]      | .452           | .667       | .499 | 1.571  | .421                               | 5.862       | .160                   | .178                       | .969                |
|   | [ed=5]      | 0 <sup>g</sup> |            |      |        |                                    |             |                        |                            |                     |
|   | address     | .015           | .011       | .157 | 1.016  | .994                               | 1.038       | .118                   | .127                       | .977                |
| Total service                                 | Intercept   | 1.086          | .412       | .009 |        |                                    |             | .058                   | .060                       | .989                |
|   | employ      | .042           | .012       | .001 | 1.043  | 1.018                              | 1.068       | .069                   | .071                       | .986                |
|   | [marital=0] | -.659          | .201       | .001 | .517   | .349                               | .767        | .090                   | .095                       | .982                |
|   | [marital=1] | 0 <sup>g</sup> |            |      |        |                                    |             |                        |                            |                     |
|   | [ed=1]      | -3.492         | .534       | .000 | .030   | .011                               | .087        | .098                   | .104                       | .981                |
|   | [ed=2]      | -1.772         | .433       | .000 | .170   | .073                               | .398        | .075                   | .078                       | .985                |
|   | [ed=3]      | -1.307         | .441       | .003 | .271   | .114                               | .643        | .064                   | .066                       | .987                |
|   | [ed=4]      | -.488          | .432       | .259 | .614   | .263                               | 1.432       | .076                   | .079                       | .985                |
|   | [ed=5]      | 0 <sup>g</sup> |            |      |        |                                    |             |                        |                            |                     |
|   | address     | .013           | .013       | .320 | 1.013  | .987                               | 1.039       | .211                   | .244                       | .960                |

This view shows all the statistics for the pooled results. You can use and interpret these coefficients in the same way you would use this table for a dataset with no missing values.

The parameter estimates table summarizes the effect of each predictor. The ratio of the coefficient to its standard error, squared, equals the Wald statistic. If the significance level of the Wald statistic is small (less than 0.05) then the parameter is different from 0.

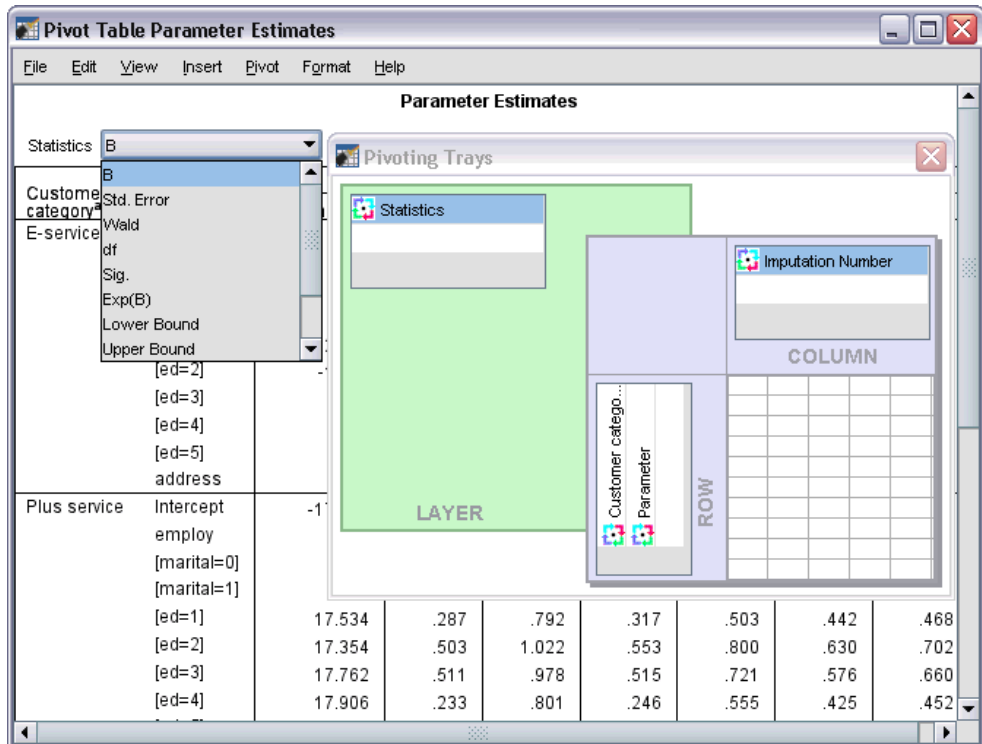
- Parameters with significant negative coefficients decrease the likelihood of that response category with respect to the reference category.

- Parameters with positive coefficients increase the likelihood of that response category.
- The parameters associated with the last category of each factor is redundant given the intercept term.

There are three additional columns in the table that provide more information about the pooled output. The **fraction of missing information** is an estimate of the ratio of missing information to “complete” information, based on the **relative increase in variance** due to non-response, which in turn is a (modified) ratio of the between-imputation and average within-imputation variance of the regression coefficient. The **relative efficiency** is a comparison of this estimate to a (theoretical) estimate computed using an infinite number of imputations. The relative efficiency is determined by the fraction of missing information and the number of imputations used to obtain the pooled result; when the fraction of missing information is large, a greater number of imputations are necessary to bring the relative efficiency closer to 1 and the pooled estimate closer to the idealized estimate.



Figure 5-38  
Pooled parameter estimates



- ▶ Now reactivate (double-click) the table, then select Pivoting Trays from the context menu.
- ▶ Move *Imputation Number* from the Layer into the Column.
- ▶ Move *Statistics* from the Column into the Layer.
- ▶ Select *B* from the Statistics dropdown list.

**Figure 5-39**  
*Pooled parameter estimates, Imputation Number in Columns and Statistics in Layer*

| Statistics= B                    |             | Imputation Number |                |                |                |                |                |                |
|----------------------------------|-------------|-------------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Customer category <sup>a,b</sup> | Parameter   | Original data     | 1              | 2              | 3              | 4              | 5              | Pooled         |
| E-service                        | Intercept   | .637              | .605           | .366           | .613           | .627           | .555           | .553           |
|                                  | employ      | .054              | .033           | .030           | .028           | .027           | .030           | .030           |
|                                  | [marital=0] | -.760             | -.497          | -.621          | -.553          | -.604          | -.551          | -.565          |
|                                  | [marital=1] | 0 <sup>a</sup>    | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> |
|                                  | [ed=1]      | -2.272            | -2.171         | -1.928         | -2.111         | -2.223         | -2.007         | -2.088         |
|                                  | [ed=2]      | -1.657            | -1.435         | -1.131         | -1.412         | -1.383         | -1.302         | -1.333         |
|                                  | [ed=3]      | -.848             | -.870          | -.584          | -.816          | -.837          | -.764          | -.774          |
|                                  | [ed=4]      | -.576             | -.610          | -.353          | -.665          | -.678          | -.491          | -.559          |
|                                  | [ed=5]      | 0 <sup>a</sup>    | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> |
|                                  | address     | .028              | .028           | .033           | .032           | .036           | .027           | .031           |
| Plus service                     | Intercept   | -17.912           | -.915          | -1.425         | -.947          | -1.191         | -1.041         | -1.104         |
|                                  | employ      | .056              | .055           | .049           | .051           | .052           | .054           | .053           |
|                                  | [marital=0] | -.549             | -.332          | -.352          | -.343          | -.330          | -.306          | -.333          |
|                                  | [marital=1] | 0 <sup>a</sup>    | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> |
|                                  | [ed=1]      | 17.534            | .287           | .792           | .317           | .503           | .442           | .468           |
|                                  | [ed=2]      | 17.354            | .503           | 1.022          | .553           | .800           | .630           | .702           |
|                                  | [ed=3]      | 17.762            | .511           | .978           | .515           | .721           | .576           | .660           |
|                                  | [ed=4]      | 17.906            | .233           | .801           | .246           | .555           | .425           | .452           |
|                                  | [ed=5]      | 0 <sup>a</sup>    | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> |
|                                  | address     | .023              | .012           | .019           | .017           | .017           | .012           | .015           |
| Total service                    | Intercept   | 1.266             | 1.157          | .942           | 1.153          | 1.118          | 1.058          | 1.086          |
|                                  | employ      | .044              | .041           | .039           | .046           | .040           | .044           | .042           |
|                                  | [marital=0] | -.522             | -.647          | -.749          | -.612          | -.666          | -.624          | -.659          |
|                                  | [marital=1] | 0 <sup>a</sup>    | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> |
|                                  | [ed=1]      | -3.590            | -3.406         | -3.390         | -3.727         | -3.555         | -3.380         | -3.492         |
|                                  | [ed=2]      | -2.133            | -1.796         | -1.594         | -1.818         | -1.878         | -1.776         | -1.772         |
|                                  | [ed=3]      | -1.214            | -1.347         | -1.146         | -1.398         | -1.366         | -1.276         | -1.307         |
|                                  | [ed=4]      | -.468             | -.548          | -.335          | -.578          | -.562          | -.417          | -.488          |
|                                  | [ed=5]      | 0 <sup>a</sup>    | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> | 0 <sup>a</sup> |
|                                  | address     | .012              | .011           | .018           | .008           | .019           | .008           | .013           |

This view of the table is useful for comparing values across imputations, to get a quick visual check of the variation in the regression coefficient estimates from imputation to imputation, and even against the original data. In particular, switching the statistic in the layer to Std. Error allows you to see how multiple imputation has reduced the variability in the coefficient estimates versus listwise deletion (original data).

Figure 5-40  
Warnings

The following variables: retire, gender, age, reside, lninc are used only to define the subpopulations but not in constructing the model.

For split file Imputation Number = Original data, unexpected singularities in the Hessian matrix are encountered. This indicates that either some predictor variables should be excluded or some categories should be merged.

The NOMREG procedure continues despite the above warning(s). Subsequent results shown are based on the last iteration. Validity of the model fit is uncertain.

However, in this example, the original dataset actually causes an error, which explains the very large parameter estimates for the *Plus service* intercept and non-redundant levels of *ed* (*Level of education*) in the original data column of the table.

## Summary

Using the multiple imputation procedures, you have analyzed patterns of missing values and found that much information would likely be lost if simple listwise deletion were used. After an initial automatic run of multiple imputation, you found that constraints were needed to keep imputed values within reasonable bounds. The run with constraints produced good values, and there was no immediate evidence that the FCS method did not converge. Using the “complete” dataset with multiply imputed values, you fit a Multinomial Logistic Regression to the data and obtained pooled regression estimates and also discovered that the final model fit would, in fact, not have been possible using listwise deletion on the original data.

# ***Sample Files***

The sample files installed with the product can be found in the *Samples* subdirectory of the installation directory. There is a separate folder within the Samples subdirectory for each of the following languages: English, French, German, Italian, Japanese, Korean, Polish, Russian, Simplified Chinese, Spanish, and Traditional Chinese.

Not all sample files are available in all languages. If a sample file is not available in a language, that language folder contains an English version of the sample file.

## ***Descriptions***

Following are brief descriptions of the sample files used in various examples throughout the documentation.

- **accidents.sav.** This is a hypothetical data file that concerns an insurance company that is studying age and gender risk factors for automobile accidents in a given region. Each case corresponds to a cross-classification of age category and gender.
- **adl.sav.** This is a hypothetical data file that concerns efforts to determine the benefits of a proposed type of therapy for stroke patients. Physicians randomly assigned female stroke patients to one of two groups. The first received the standard physical therapy, and the second received an additional emotional therapy. Three months following the treatments, each patient's abilities to perform common activities of daily life were scored as ordinal variables.
- **advert.sav.** This is a hypothetical data file that concerns a retailer's efforts to examine the relationship between money spent on advertising and the resulting sales. To this end, they have collected past sales figures and the associated advertising costs..

- **aflatoxin.sav.** This is a hypothetical data file that concerns the testing of corn crops for aflatoxin, a poison whose concentration varies widely between and within crop yields. A grain processor has received 16 samples from each of 8 crop yields and measured the aflatoxin levels in parts per billion (PPB).
- **aflatoxin20.sav.** This data file contains the aflatoxin measurements from each of the 16 samples from yields 4 and 8 from the *aflatoxin.sav* data file.
- **anorectic.sav.** While working toward a standardized symptomatology of anorectic/bulimic behavior, researchers made a study of 55 adolescents with known eating disorders. Each patient was seen four times over four years, for a total of 220 observations. At each observation, the patients were scored for each of 16 symptoms. Symptom scores are missing for patient 71 at time 2, patient 76 at time 2, and patient 47 at time 3, leaving 217 valid observations.
- **autoaccidents.sav.** This is a hypothetical data file that concerns the efforts of an insurance analyst to model the number of automobile accidents per driver while also accounting for driver age and gender. Each case represents a separate driver and records the driver's gender, age in years, and number of automobile accidents in the last five years.
- **band.sav.** This data file contains hypothetical weekly sales figures of music CDs for a band. Data for three possible predictor variables are also included.
- **bankloan.sav.** This is a hypothetical data file that concerns a bank's efforts to reduce the rate of loan defaults. The file contains financial and demographic information on 850 past and prospective customers. The first 700 cases are customers who were previously given loans. The last 150 cases are prospective customers that the bank needs to classify as good or bad credit risks.
- **bankloan\_binning.sav.** This is a hypothetical data file containing financial and demographic information on 5,000 past customers.
- **behavior.sav.** In a classic example, 52 students were asked to rate the combinations of 15 situations and 15 behaviors on a 10-point scale ranging from 0="extremely appropriate" to 9="extremely inappropriate." Averaged over individuals, the values are taken as dissimilarities.
- **behavior\_ini.sav.** This data file contains an initial configuration for a two-dimensional solution for *behavior.sav*.
- **brakes.sav.** This is a hypothetical data file that concerns quality control at a factory that produces disc brakes for high-performance automobiles. The data file contains diameter measurements of 16 discs from each of 8 production machines. The target diameter for the brakes is 322 millimeters.

- **breakfast.sav.** In a classic study, 21 Wharton School MBA students and their spouses were asked to rank 15 breakfast items in order of preference with 1=“most preferred” to 15=“least preferred.” Their preferences were recorded under six different scenarios, from “Overall preference” to “Snack, with beverage only.”
- **breakfast-overall.sav.** This data file contains the breakfast item preferences for the first scenario, “Overall preference,” only.
- **broadband\_1.sav.** This is a hypothetical data file containing the number of subscribers, by region, to a national broadband service. The data file contains monthly subscriber numbers for 85 regions over a four-year period.
- **broadband\_2.sav.** This data file is identical to *broadband\_1.sav* but contains data for three additional months.
- **car\_insurance\_claims.sav.** A dataset presented and analyzed elsewhere concerns damage claims for cars. The average claim amount can be modeled as having a gamma distribution, using an inverse link function to relate the mean of the dependent variable to a linear combination of the policyholder age, vehicle type, and vehicle age. The number of claims filed can be used as a scaling weight.
- **car\_sales.sav.** This data file contains hypothetical sales estimates, list prices, and physical specifications for various makes and models of vehicles. The list prices and physical specifications were obtained alternately from *edmunds.com* and manufacturer sites.
- **carpet.sav.** In a popular example, a company interested in marketing a new carpet cleaner wants to examine the influence of five factors on consumer preference—package design, brand name, price, a *Good Housekeeping* seal, and a money-back guarantee. There are three factor levels for package design, each one differing in the location of the applicator brush; three brand names (*K2R*, *Glory*, and *Bissell*); three price levels; and two levels (either no or yes) for each of the last two factors. Ten consumers rank 22 profiles defined by these factors. The variable *Preference* contains the rank of the average rankings for each profile. Low rankings correspond to high preference. This variable reflects an overall measure of preference for each profile.
- **carpet\_prefs.sav.** This data file is based on the same example as described for *carpet.sav*, but it contains the actual rankings collected from each of the 10 consumers. The consumers were asked to rank the 22 product profiles from the most to the least preferred. The variables *PREF1* through *PREF22* contain the identifiers of the associated profiles, as defined in *carpet\_plan.sav*.

- 
- **catalog.sav.** This data file contains hypothetical monthly sales figures for three products sold by a catalog company. Data for five possible predictor variables are also included.
  - **catalog\_seasfac.sav.** This data file is the same as *catalog.sav* except for the addition of a set of seasonal factors calculated from the Seasonal Decomposition procedure along with the accompanying date variables.
  - **cellular.sav.** This is a hypothetical data file that concerns a cellular phone company's efforts to reduce churn. Churn propensity scores are applied to accounts, ranging from 0 to 100. Accounts scoring 50 or above may be looking to change providers.
  - **ceramics.sav.** This is a hypothetical data file that concerns a manufacturer's efforts to determine whether a new premium alloy has a greater heat resistance than a standard alloy. Each case represents a separate test of one of the alloys; the heat at which the bearing failed is recorded.
  - **cereal.sav.** This is a hypothetical data file that concerns a poll of 880 people about their breakfast preferences, also noting their age, gender, marital status, and whether or not they have an active lifestyle (based on whether they exercise at least twice a week). Each case represents a separate respondent.
  - **clothing\_defects.sav.** This is a hypothetical data file that concerns the quality control process at a clothing factory. From each lot produced at the factory, the inspectors take a sample of clothes and count the number of clothes that are unacceptable.
  - **coffee.sav.** This data file pertains to perceived images of six iced-coffee brands . For each of 23 iced-coffee image attributes, people selected all brands that were described by the attribute. The six brands are denoted AA, BB, CC, DD, EE, and FF to preserve confidentiality.
  - **contacts.sav.** This is a hypothetical data file that concerns the contact lists for a group of corporate computer sales representatives. Each contact is categorized by the department of the company in which they work and their company ranks. Also recorded are the amount of the last sale made, the time since the last sale, and the size of the contact's company.
  - **creditpromo.sav.** This is a hypothetical data file that concerns a department store's efforts to evaluate the effectiveness of a recent credit card promotion. To this end, 500 cardholders were randomly selected. Half received an ad promoting a reduced interest rate on purchases made over the next three months. Half received a standard seasonal ad.

- **customer\_dbase.sav.** This is a hypothetical data file that concerns a company's efforts to use the information in its data warehouse to make special offers to customers who are most likely to reply. A subset of the customer base was selected at random and given the special offers, and their responses were recorded.
- **customer\_information.sav.** A hypothetical data file containing customer mailing information, such as name and address.
- **customers\_model.sav.** This file contains hypothetical data on individuals targeted by a marketing campaign. These data include demographic information, a summary of purchasing history, and whether or not each individual responded to the campaign. Each case represents a separate individual.
- **customers\_new.sav.** This file contains hypothetical data on individuals who are potential candidates for a marketing campaign. These data include demographic information and a summary of purchasing history for each individual. Each case represents a separate individual.
- **debate.sav.** This is a hypothetical data file that concerns paired responses to a survey from attendees of a political debate before and after the debate. Each case corresponds to a separate respondent.
- **debate\_aggregate.sav.** This is a hypothetical data file that aggregates the responses in *debate.sav*. Each case corresponds to a cross-classification of preference before and after the debate.
- **demo.sav.** This is a hypothetical data file that concerns a purchased customer database, for the purpose of mailing monthly offers. Whether or not the customer responded to the offer is recorded, along with various demographic information.
- **demo\_cs\_1.sav.** This is a hypothetical data file that concerns the first step of a company's efforts to compile a database of survey information. Each case corresponds to a different city, and the region, province, district, and city identification are recorded.
- **demo\_cs\_2.sav.** This is a hypothetical data file that concerns the second step of a company's efforts to compile a database of survey information. Each case corresponds to a different household unit from cities selected in the first step, and the region, province, district, city, subdivision, and unit identification are recorded. The sampling information from the first two stages of the design is also included.
- **demo\_cs.sav.** This is a hypothetical data file that contains survey information collected using a complex sampling design. Each case corresponds to a different household unit, and various demographic and sampling information is recorded.



- **dietstudy.sav.** This hypothetical data file contains the results of a study of the “Stillman diet” . Each case corresponds to a separate subject and records his or her pre- and post-diet weights in pounds and triglyceride levels in mg/100 ml.
- **dischargedata.sav.** This is a data file concerning *Seasonal Patterns of Winnipeg Hospital Use*, from the Manitoba Centre for Health Policy.
- **dvdplayer.sav.** This is a hypothetical data file that concerns the development of a new DVD player. Using a prototype, the marketing team has collected focus group data. Each case corresponds to a separate surveyed user and records some demographic information about them and their responses to questions about the prototype.
- **flying.sav.** This data file contains the flying mileages between 10 American cities.
- **german\_credit.sav.** This data file is taken from the “German credit” dataset in the Repository of Machine Learning Databases at the University of California, Irvine.
- **grocery\_1month.sav.** This hypothetical data file is the *grocery\_coupons.sav* data file with the weekly purchases “rolled-up” so that each case corresponds to a separate customer. Some of the variables that changed weekly disappear as a result, and the amount spent recorded is now the sum of the amounts spent during the four weeks of the study.
- **grocery\_coupons.sav.** This is a hypothetical data file that contains survey data collected by a grocery store chain interested in the purchasing habits of their customers. Each customer is followed for four weeks, and each case corresponds to a separate customer-week and records information about where and how the customer shops, including how much was spent on groceries during that week.
- **guttman.sav.** Bell presented a table to illustrate possible social groups. Guttman used a portion of this table, in which five variables describing such things as social interaction, feelings of belonging to a group, physical proximity of members, and formality of the relationship were crossed with seven theoretical social groups, including crowds (for example, people at a football game), audiences (for example, people at a theater or classroom lecture), public (for example, newspaper or television audiences), mobs (like a crowd but with much more intense interaction), primary groups (intimate), secondary groups (voluntary), and the modern community (loose confederation resulting from close physical proximity and a need for specialized services).
- **healthplans.sav.** This is a hypothetical data file that concerns an insurance group’s efforts to evaluate four different health care plans for small employers. Twelve employers are recruited to rank the plans by how much they would prefer to

offer them to their employees. Each case corresponds to a separate employer and records the reactions to each plan.

- **health\_funding.sav.** This is a hypothetical data file that contains data on health care funding (amount per 100 population), disease rates (rate per 10,000 population), and visits to health care providers (rate per 10,000 population). Each case represents a different city.
- **hivassay.sav.** This is a hypothetical data file that concerns the efforts of a pharmaceutical lab to develop a rapid assay for detecting HIV infection. The results of the assay are eight deepening shades of red, with deeper shades indicating greater likelihood of infection. A laboratory trial was conducted on 2,000 blood samples, half of which were infected with HIV and half of which were clean.
- **hourlywagedata.sav.** This is a hypothetical data file that concerns the hourly wages of nurses from office and hospital positions and with varying levels of experience.
- **insure.sav.** This is a hypothetical data file that concerns an insurance company that is studying the risk factors that indicate whether a client will have to make a claim on a 10-year term life insurance contract. Each case in the data file represents a pair of contracts, one of which recorded a claim and the other didn't, matched on age and gender.
- **judges.sav.** This is a hypothetical data file that concerns the scores given by trained judges (plus one enthusiast) to 300 gymnastics performances. Each row represents a separate performance; the judges viewed the same performances.
- **kinship\_dat.sav.** Rosenberg and Kim set out to analyze 15 kinship terms (aunt, brother, cousin, daughter, father, granddaughter, grandfather, grandmother, grandson, mother, nephew, niece, sister, son, uncle). They asked four groups of college students (two female, two male) to sort these terms on the basis of similarities. Two groups (one female, one male) were asked to sort twice, with the second sorting based on a different criterion from the first sort. Thus, a total of six "sources" were obtained. Each source corresponds to a  $15 \times 15$  proximity matrix, whose cells are equal to the number of people in a source minus the number of times the objects were partitioned together in that source.
- **kinship\_ini.sav.** This data file contains an initial configuration for a three-dimensional solution for *kinship\_dat.sav*.
- **kinship\_var.sav.** This data file contains independent variables *gender*, *gener(ation)*, and *degree* (of separation) that can be used to interpret the dimensions of a solution for *kinship\_dat.sav*. Specifically, they can be used to restrict the space of the solution to a linear combination of these variables.

- **mailresponse.sav.** This is a hypothetical data file that concerns the efforts of a clothing manufacturer to determine whether using first class postage for direct mailings results in faster responses than bulk mail. Order-takers record how many weeks after the mailing each order is taken.
- **marketvalues.sav.** This data file concerns home sales in a new housing development in Algonquin, Ill., during the years from 1999–2000. These sales are a matter of public record.
- **mutualfund.sav.** This data file concerns stock market information for various tech stocks listed on the S&P 500. Each case corresponds to a separate company.
- **nhis2000\_subset.sav.** The National Health Interview Survey (NHIS) is a large, population-based survey of the U.S. civilian population. Interviews are carried out face-to-face in a nationally representative sample of households. Demographic information and observations about health behaviors and status are obtained for members of each household. This data file contains a subset of information from the 2000 survey. National Center for Health Statistics. National Health Interview Survey, 2000. Public-use data file and documentation. [ftp://ftp.cdc.gov/pub/Health\\_Statistics/NCHS/Datasets/NHIS/2000/](ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/). Accessed 2003.
- **ozone.sav.** The data include 330 observations on six meteorological variables for predicting ozone concentration from the remaining variables. Previous researchers , , among others found nonlinearities among these variables, which hinder standard regression approaches.
- **pain\_medication.sav.** This hypothetical data file contains the results of a clinical trial for anti-inflammatory medication for treating chronic arthritic pain. Of particular interest is the time it takes for the drug to take effect and how it compares to an existing medication.
- **patient\_los.sav.** This hypothetical data file contains the treatment records of patients who were admitted to the hospital for suspected myocardial infarction (MI, or “heart attack”). Each case corresponds to a separate patient and records many variables related to their hospital stay.
- **patlos\_sample.sav.** This hypothetical data file contains the treatment records of a sample of patients who received thrombolytics during treatment for myocardial infarction (MI, or “heart attack”). Each case corresponds to a separate patient and records many variables related to their hospital stay.

- **polishing.sav.** This is the “Nambeware Polishing Times” data file from the Data and Story Library. It concerns the efforts of a metal tableware manufacturer (Nambe Mills, Santa Fe, N. M.) to plan its production schedule. Each case represents a different item in the product line. The diameter, polishing time, price, and product type are recorded for each item.
- **poll\_cs.sav.** This is a hypothetical data file that concerns pollsters’ efforts to determine the level of public support for a bill before the legislature. The cases correspond to registered voters. Each case records the county, township, and neighborhood in which the voter lives.
- **poll\_cs\_sample.sav.** This hypothetical data file contains a sample of the voters listed in *poll\_cs.sav*. The sample was taken according to the design specified in the *poll.csplan* plan file, and this data file records the inclusion probabilities and sample weights. Note, however, that because the sampling plan makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*poll\_jointprob.sav*). The additional variables corresponding to voter demographics and their opinion on the proposed bill were collected and added to the data file after the sample was taken.
- **property\_assess.sav.** This is a hypothetical data file that concerns a county assessor’s efforts to keep property value assessments up to date on limited resources. The cases correspond to properties sold in the county in the past year. Each case in the data file records the township in which the property lies, the assessor who last visited the property, the time since that assessment, the valuation made at that time, and the sale value of the property.
- **property\_assess\_cs.sav.** This is a hypothetical data file that concerns a state assessor’s efforts to keep property value assessments up to date on limited resources. The cases correspond to properties in the state. Each case in the data file records the county, township, and neighborhood in which the property lies, the time since the last assessment, and the valuation made at that time.
- **property\_assess\_cs\_sample.sav.** This hypothetical data file contains a sample of the properties listed in *property\_assess\_cs.sav*. The sample was taken according to the design specified in the *property\_assess.csplan* plan file, and this data file records the inclusion probabilities and sample weights. The additional variable *Current value* was collected and added to the data file after the sample was taken.
- **recidivism.sav.** This is a hypothetical data file that concerns a government law enforcement agency’s efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender and records their

demographic information, some details of their first crime, and then the time until their second arrest, if it occurred within two years of the first arrest.

- **recidivism\_cs\_sample.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender, released from their first arrest during the month of June, 2003, and records their demographic information, some details of their first crime, and the data of their second arrest, if it occurred by the end of June, 2006. Offenders were selected from sampled departments according to the sampling plan specified in *recidivism\_cs.csplan*; because it makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*recidivism\_cs\_jointprob.sav*).
- **rfm\_transactions.sav.** A hypothetical data file containing purchase transaction data, including date of purchase, item(s) purchased, and monetary amount of each transaction.
- **salesperformance.sav.** This is a hypothetical data file that concerns the evaluation of two new sales training courses. Sixty employees, divided into three groups, all receive standard training. In addition, group 2 gets technical training; group 3, a hands-on tutorial. Each employee was tested at the end of the training course and their score recorded. Each case in the data file represents a separate trainee and records the group to which they were assigned and the score they received on the exam.
- **satisf.sav.** This is a hypothetical data file that concerns a satisfaction survey conducted by a retail company at 4 store locations. 582 customers were surveyed in all, and each case represents the responses from a single customer.
- **screws.sav.** This data file contains information on the characteristics of screws, bolts, nuts, and tacks .
- **shampoo\_ph.sav.** This is a hypothetical data file that concerns the quality control at a factory for hair products. At regular time intervals, six separate output batches are measured and their pH recorded. The target range is 4.5–5.5.
- **ships.sav.** A dataset presented and analyzed elsewhere that concerns damage to cargo ships caused by waves. The incident counts can be modeled as occurring at a Poisson rate given the ship type, construction period, and service period. The aggregate months of service for each cell of the table formed by the cross-classification of factors provides values for the exposure to risk.

- **site.sav.** This is a hypothetical data file that concerns a company's efforts to choose new sites for their expanding business. They have hired two consultants to separately evaluate the sites, who, in addition to an extended report, summarized each site as a "good," "fair," or "poor" prospect.
- **siteratings.sav.** This is a hypothetical data file that concerns the beta testing of an e-commerce firm's new Web site. Each case represents a separate beta tester, who scored the usability of the site on a scale from 0–20.
- **smokers.sav.** This data file is abstracted from the 1998 National Household Survey of Drug Abuse and is a probability sample of American households. Thus, the first step in an analysis of this data file should be to weight the data to reflect population trends.
- **smoking.sav.** This is a hypothetical table introduced by Greenacre . The table of interest is formed by the crosstabulation of smoking behavior by job category. The variable *Staff Group* contains the job categories *Sr Managers*, *Jr Managers*, *Sr Employees*, *Jr Employees*, and *Secretaries*, plus the category *National Average*, which can be used as supplementary to an analysis. The variable *Smoking* contains the behaviors *None*, *Light*, *Medium*, and *Heavy*, plus the categories *No Alcohol* and *Alcohol*, which can be used as supplementary to an analysis.
- **storebrand.sav.** This is a hypothetical data file that concerns a grocery store manager's efforts to increase sales of the store brand detergent relative to other brands. She puts together an in-store promotion and talks with customers at check-out. Each case represents a separate customer.
- **stores.sav.** This data file contains hypothetical monthly market share data for two competing grocery stores. Each case represents the market share data for a given month.
- **stroke\_clean.sav.** This hypothetical data file contains the state of a medical database after it has been cleaned using procedures in the Data Preparation option.
- **stroke\_invalid.sav.** This hypothetical data file contains the initial state of a medical database and contains several data entry errors.
- **stroke\_survival.** This hypothetical data file concerns survival times for patients exiting a rehabilitation program post-ischemic stroke face a number of challenges. Post-stroke, the occurrence of myocardial infarction, ischemic stroke, or hemorrhagic stroke is noted and the time of the event recorded. The sample is left-truncated because it only includes patients who survived through the end of the rehabilitation program administered post-stroke.

- 
- **stroke\_valid.sav.** This hypothetical data file contains the state of a medical database after the values have been checked using the Validate Data procedure. It still contains potentially anomalous cases.
  - **survey\_sample.sav.** This hypothetical data file contains survey data, including demographic data and various attitude measures.
  - **tastetest.sav.** This is a hypothetical data file that concerns the effect of mulch color on the taste of crops. Strawberries grown in red, blue, and black mulch were rated by taste-testers on an ordinal scale of 1 to 5 (far below to far above average). Each case represents a separate taste-tester.
  - **telco.sav.** This is a hypothetical data file that concerns a telecommunications company's efforts to reduce churn in their customer base. Each case corresponds to a separate customer and records various demographic and service usage information.
  - **telco\_extra.sav.** This data file is similar to the *telco.sav* data file, but the "tenure" and log-transformed customer spending variables have been removed and replaced by standardized log-transformed customer spending variables.
  - **telco\_missing.sav.** This data file is a subset of the *telco.sav* data file, but some of the demographic data values have been replaced with missing values.
  - **testmarket.sav.** This hypothetical data file concerns a fast food chain's plans to add a new item to its menu. There are three possible campaigns for promoting the new product, so the new item is introduced at locations in several randomly selected markets. A different promotion is used at each location, and the weekly sales of the new item are recorded for the first four weeks. Each case corresponds to a separate location-week.
  - **testmarket\_1month.sav.** This hypothetical data file is the *testmarket.sav* data file with the weekly sales "rolled-up" so that each case corresponds to a separate location. Some of the variables that changed weekly disappear as a result, and the sales recorded is now the sum of the sales during the four weeks of the study.
  - **tree\_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.
  - **tree\_credit.sav.** This is a hypothetical data file containing demographic and bank loan history data.
  - **tree\_missing\_data.sav** This is a hypothetical data file containing demographic and bank loan history data with a large number of missing values.

- **tree\_score\_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.
- **tree\_textdata.sav.** A simple data file with only two variables intended primarily to show the default state of variables prior to assignment of measurement level and value labels.
- **tv-survey.sav.** This is a hypothetical data file that concerns a survey conducted by a TV studio that is considering whether to extend the run of a successful program. 906 respondents were asked whether they would watch the program under various conditions. Each row represents a separate respondent; each column is a separate condition.
- **ulcer\_recurrence.sav.** This file contains partial information from a study designed to compare the efficacy of two therapies for preventing the recurrence of ulcers. It provides a good example of interval-censored data and has been presented and analyzed elsewhere .
- **ulcer\_recurrence\_recoded.sav.** This file reorganizes the information in *ulcer\_recurrence.sav* to allow you model the event probability for each interval of the study rather than simply the end-of-study event probability. It has been presented and analyzed elsewhere .
- **verd1985.sav.** This data file concerns a survey . The responses of 15 subjects to 8 variables were recorded. The variables of interest are divided into three sets. Set 1 includes *age* and *marital*, set 2 includes *pet* and *news*, and set 3 includes *music* and *live*. *Pet* is scaled as multiple nominal and *age* is scaled as ordinal; all of the other variables are scaled as single nominal.
- **virus.sav.** This is a hypothetical data file that concerns the efforts of an Internet service provider (ISP) to determine the effects of a virus on its networks. They have tracked the (approximate) percentage of infected e-mail traffic on its networks over time, from the moment of discovery until the threat was contained.
- **waittimes.sav.** This is a hypothetical data file that concerns customer waiting times for service at three different branches of a local bank. Each case corresponds to a separate customer and records the time spent waiting and the branch at which they were conducting their business.
- **webusability.sav.** This is a hypothetical data file that concerns usability testing of a new e-store. Each case corresponds to one of five usability testers and records whether or not the tester succeeded at each of six separate tasks.



- **wheeze\_steubenville.sav.** This is a subset from a longitudinal study of the health effects of air pollution on children . The data contain repeated binary measures of the wheezing status for children from Steubenville, Ohio, at ages 7, 8, 9 and 10 years, along with a fixed recording of whether or not the mother was a smoker during the first year of the study.
- **workprog.sav.** This is a hypothetical data file that concerns a government works program that tries to place disadvantaged people into better jobs. A sample of potential program participants were followed, some of whom were randomly selected for enrollment in the program, while others were not. Each case represents a separate program participant.

- Analyze Patterns, 19
- correlations
  - in Missing Value Analysis, 11, 13
- covariance
  - in Missing Value Analysis, 11, 13
- EM
  - in Missing Value Analysis, 11
- extreme value counts
  - in Missing Value Analysis, 8
- FCS convergence chart
  - in multiple imputation, 86
- frequency tables
  - in Missing Value Analysis, 8
- fully conditional specification
  - in Multiple Imputation, 24
- Impute Missing Data Values, 21
  - constraints, 27
  - imputation method, 24
  - output, 30
- incomplete data
  - see Missing Value Analysis, 3
- indicator variables
  - in Missing Value Analysis, 8
- iteration history
  - in Multiple Imputation, 30
- listwise deletion
  - in Missing Value Analysis, 3
- Little's MCAR test, 10
  - in Missing Value Analysis, 3, 57
- MCAR test
  - in Missing Value Analysis, 3, 57
- mean
  - in Missing Value Analysis, 8, 11, 13
- mismatch
  - in Missing Value Analysis, 8
- missing indicator variables
  - in Missing Value Analysis, 8
- Missing Value Analysis, 3, 45
  - command additional features, 16
  - descriptive statistics, 8, 45
  - EM, 11
  - estimating statistics, 10
  - expectation-maximization, 15
  - imputing missing values, 10
  - MCAR test, 10
  - methods, 10
  - patterns, 6, 54
  - regression, 13
- missing value patterns, 56
- missing values
  - univariate statistics, 8, 48
- monotone imputation
  - in Multiple Imputation, 24
- multiple imputation, 17, 59
  - analyze patterns, 19
  - constraints, 80
  - descriptive statistics, 70, 80
  - FCS convergence chart, 86
  - imputation results, 68
  - imputation specifications, 68
  - impute missing data values, 21
  - missing value patterns, 63
  - models, 69
  - overall summary of missing values, 61
  - pooled estimates, 93
  - pooled results, 87
  - variable summary, 62
- Multiple Imputation, 31, 36
  - options, 42

- 
- normal variates
    - in Missing Value Analysis, 13
  
  - options
    - multiple imputation, 42
  
  - pairwise deletion
    - in Missing Value Analysis, 3
  - pooled estimates
    - in multiple imputation, 93
  - pooled results
    - in multiple imputation, 87
  
  - regression
    - in Missing Value Analysis, 13
  - residuals
    - in Missing Value Analysis, 13
  
  - sample files
    - location, 100
  - sorting cases
    - in Missing Value Analysis, 6
  - standard deviation
    - in Missing Value Analysis, 8
  - Student's t test
    - in Missing Value Analysis, 13, 49
  
  - t test
    - in Missing Value Analysis, 8, 49
  - tabulating cases
    - in Missing Value Analysis, 6
  - tabulating categories
    - in Missing Value Analysis, 8, 50
  
  - univariate statistics
    - in Missing Value Analysis, 48