SOLAS™ Version 4.0

MISSING DATA

SINGLE IMPUTATION

MULTIPLE IMPUTATION

ANALYZING MULTIPLY-IMPUTED DATASETS

APPENDICES

About this Manual

This manual deals with the problem of analyzing data sets in which data are missing. We first explain how to run SOLAS™ so you can begin your analyses. We then provide some general information about handling variables and cases in SOLAS™, and this is followed by overviews of the Single and Multiple Imputation techniques that are available in the new SOLAS™ 4.0, followed by examples using Single Imputation.

Multiple Imputation is then discussed in more detail with a description of the method SOLAS™ uses to sort Monotone and Non-monotone missing data and displays the data patterns. Then each of the five Multiple Imputation techniques are described, and a description of the way in which SOLAS™ imputes Monotone and Non-monotone missing data is given. This is followed by short examples for each of the Multiple Imputation methods. These examples demonstrate how each of the available methods for Single and Multiple Imputation are used.

Details about the output from running an imputation are given with an example of how to analyze a multiply imputed dataset.

Finally, several appendices are given that detail formulae and methods, and give references to literature.

NOTE: This manual is intended as a user reference for SOLAS™ and as a guide to using the various distinct methods of imputation that SOLAS™ 4.0 provides. It is not meant as a textbook for missing data, nor is it intended as a comprehensive description of multiple imputation. For this, the user should consult the references given in Appendix G.

Enhancements to SOLAS™ Version 4.0

The following is a list of enhancements in the new SOLAS™ Version 4.0:

- A new Mahalanobis Distance Based Multiple Imputation technique has been added (p. 35).
- A new Predictive Mean Matching Imputation technique has been added to the system (p. 39).
- A new Propensity Score/ Predictive Mean Matching/ Mahalanobis Distance Based Multiple Imputation technique has been added (p. 43).
- A new **Collapse** function to allow ease of interpretation for the Missing Data pattern (p. 18).
- New **Margin Plot** available to plot variables that contain missing values but still give some information about the cases that have missing values (p. 27).
- New **Scatterplot** available for variables with imputed values. This plot can include all (multiply) imputed values (p. 50).
- SOLAS™ 4.0 is available as a 32-bit and 64-bit application.

Descriptions of all of these enhancements are included in this manual.

Contents

Missing Data

Missing data are a pervasive problem in data analysis. Missing values lead to less efficient estimates because of the reduced size of the database, also standard complete-data methods of analysis no longer apply. For example, analyses such as multiple regression use only cases that have complete data, so including a variable with numerous missing values would severely reduce the sample size.

When cases are deleted if one or more variables have missing values, the number of remaining cases can be small even if the missing data rate is small for each variable. For example, suppose your data set has 5 variables measured at the start of study and monthly for six months. You have been told, with great pride, that each variable is 95% complete. If each of these 5 variables has a random 5% of the values missing, then the proportion of cases that are expected to be complete are 1-(.95)^35=0.834. That is, only 17% of the cases would be complete and you would lose 83% of your data.

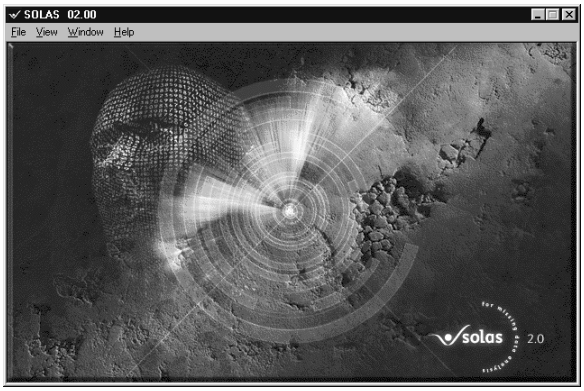
Missing data also cause difficulties in performing Intent-to-Treat analyses in randomized experiments. Intent-to-Treat (IT) analysis dictates that all cases - complete and incomplete, be included in any analyses. Biases may exist from the analysis of only complete cases if there are systematic differences between completers and dropouts. To select a valid approach for imputing missing data values for any particular variable, it is necessary to consider the underlying mechanism accounting for missing data. Variables in a data set may have values that are missing for different reasons.

A laboratory value might be missing because:

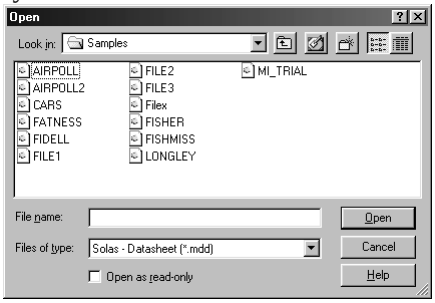
- It was below the level of detection.
- The assay was not done because the patient did not come in for a scheduled visit.
- The assay was not done because the test tube was dropped or lost.
- The assay was not done because the patient died, or was lost to follow-up, or other possible causes.

Getting Started

After performing the Setup described earlier in this manual, clicking on the SOLAS™ 4.0 icon displays the Main window shown below:

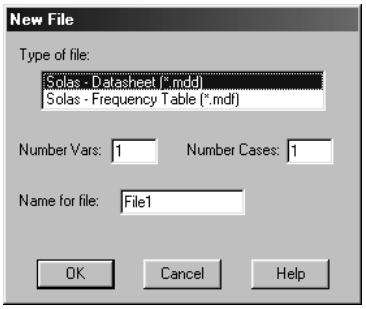


Select **File** and then **Open** from the Main window menu-bar displays an Open window. In this window you can browse the directories/folders on your system for a list of the stored data sets which you want to analyze.



The datasheets in the “Samples” folder shown in the Open window above can be used as data to perform some

example analyses which will familiarize you with the system. Several of these examples are discussed later in this manual. Alternatively, you may want to create a new datasheet, in this case you would select **New** from the **File** menu in the Main window.

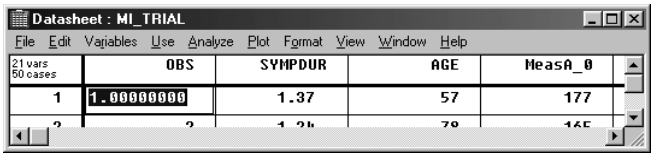


You can also set preferences for your output options from the Main window **View** menu **System Preferences** menu.

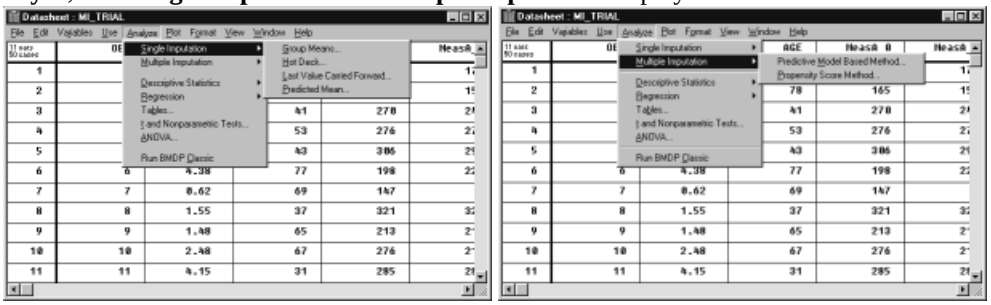
Using the data fields in this window you can create a datasheet or a frequency table with the required number of variables and cases, or rows and columns. Start in one of the following ways:

- Enter the criteria for your new datasheet, then press the **OK** button. Or
- Select an existing datasheet from your file system using the Open window, then press the **Open** button.

Whether you create a new datasheet or open an existing datasheet, you will see a window similar to the window shown below with its menu bar.



Selecting **Analyze**, then **Single Imputation** or **Multiple Imputation** displays one of the menus shown below.



From one of the menus shown above, you can select the method of imputation that you want to use, and a specification window will be displayed where the selected method can be setup.

General

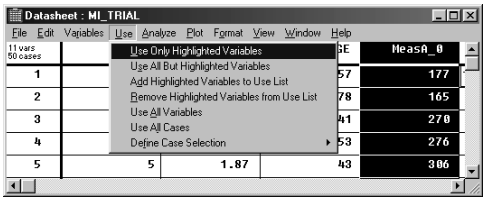
The following subsections provide general information about variables, (grouping, variable selection/de-selection, and defining case selection.

Grouping variables can be selected for all of the imputation methods. If a grouping variable is specified, then the sorting of missing data patterns and the generation of multiple imputations is carried out for each group of cases having the same observed value as the specified grouping variable.

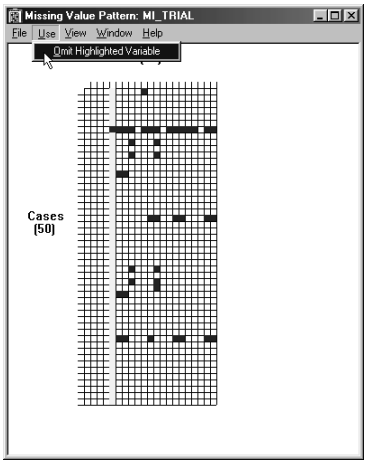
More detailed information about variables is given in Chapter 1 of the Systems Manual – Data Management, and the sections "Specifying Variable Attributes", and "Defining Variables."

Variable Selection/De-selection

There are several options regarding the variables of a data set that is to be analyzed. These options can be displayed from the datasheet **Use** menu as shown below:

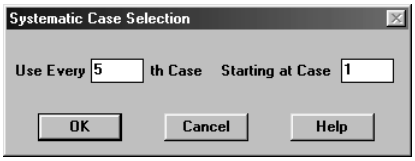


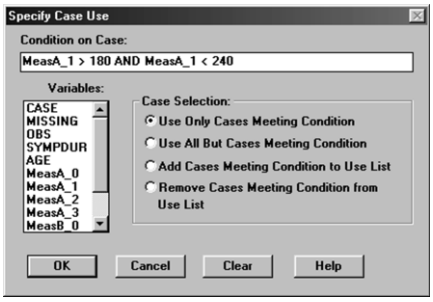
You can select and de-select variables by using the datasheet **View** menu and selecting **Missing Data Pattern** (described later) to display the Missing Data Pattern window shown below. To de-select a variable, right-click at the top of any column in the missing pattern to highlight the variable, then choose **Omit Highlighted Variable** from the **Use** menu.



Define case selection

Case selection can be applied in two ways: **Systematic** or **User-defined** by choosing **Define Case Selection** from the **Use** menu in a datasheet. Depending on the selection, one of the windows shown below is displayed:

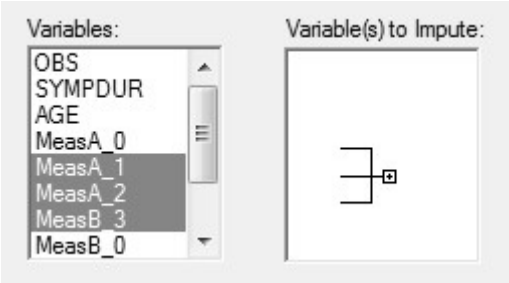




For **Systematic** case definition, numerical selection can be applied. For **User-defined** case specification, conditional and logical operators can be applied to selected cases within variables as shown in the right-hand window above. A table showing the operators, their meanings, and their keyboard entries is given in Chapter 1 of the System Manual – Data Management.

Multiple Drag and Drop

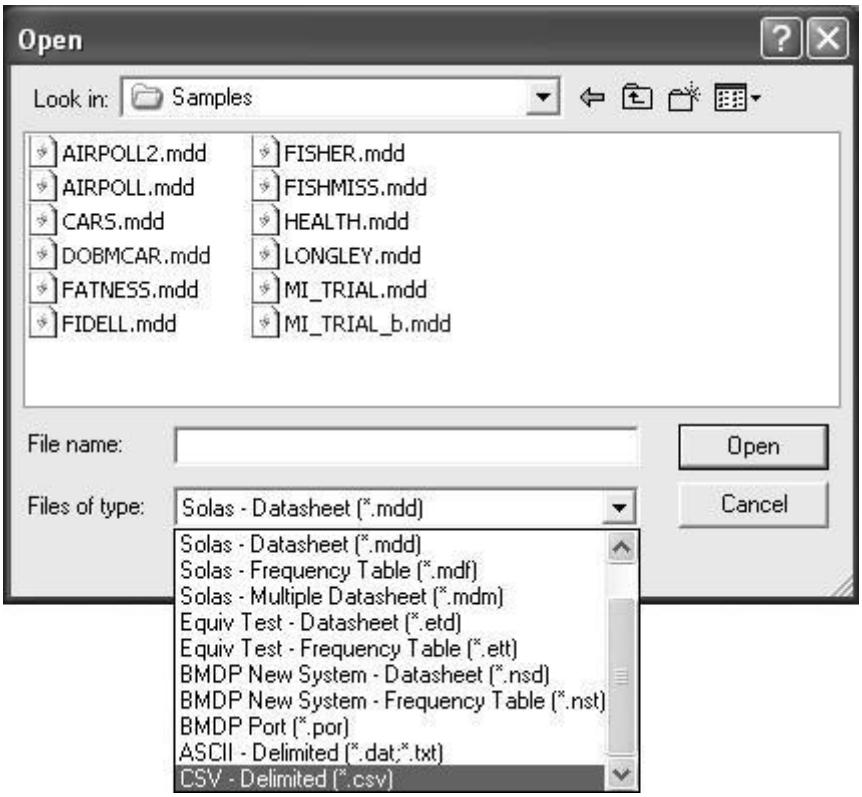
Multiple variables can be "dragged" by holding down the <Ctrl> key and selecting (highlighting) variables. You have to press the <Ctrl> key before you start selecting variables. The "Drag Variable" controls will not be enabled. If some of the variables being dragged into a data field are inappropriate for that data field, the system will display a message, and those variables will not be placed in the field. The remaining variables in the multiple selection will be moved as intended.



Opening files from other software

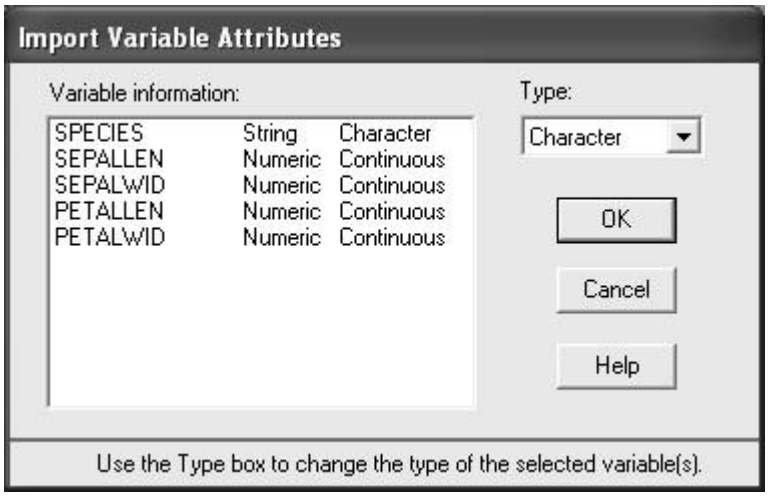
The suggested file format for bringing datasets into SOLAS is to use the comma separated variable (.csv) file format. This format is supported by most software packages and it is possible to save or export datasets as .csv.

1. Select File > Open… and choose .CSV delimited from the Files of type dropdown box.



2. Select the file to be opened and press Open.

3. Press OK. A list of the variable names and variable types will be displayed. You can use the dropdown box on the left to specify whether variables should be read as character (String) or continuous (Numeric).



4. Press OK and the dataset will open in SOLAS.

Datasheet : FISHMISS

File Edit Variables Use Analyze Plot Format View Window Help

5 vars 150 cases	SPECIES	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	Setosa	5.00	3.30	1.40	0.20
2	Virginica	6.40	2.80	5.60	2.20
3	Versicolor	6.50	2.80	4.60	1.50
4	Virginica	6.70	3.10	5.60	2.40
5	Virginica	6.30	2.80	5.10	1.50
6	Setosa	4.60	3.40	1.40	0.30
7	Virginica	6.90	3.10		2.30
8	Versicolor	6.20	2.20	4.50	1.50
9	Versicolor	5.90	3.20	4.80	1.80
10	Setosa	4.60	3.60	1.00	0.20
11	Versicolor	6.10	3.00	4.60	1.40
12	Versicolor	6.00	2.70	5.10	1.60
13	Virginica	6.50	3.00	5.20	2.00
14	Versicolor	5.60			1.10

Overviews of Imputation in SOLAS™

Imputation is the name given to any method whereby missing values in a data set are filled-in with plausible estimates. The goal of any imputation technique is to produce a complete data set that can be analyzed using complete-data inferential methods. The following describes the Single and Multiple imputation methods available in SOLAS™ 4.0 that are designed to accommodate a range of missing data scenarios in both longitudinal and single-observation study designs.

Single Imputation Overview

SOLAS™ 4.0 provides four distinct methods by which you can perform Single Imputation – Group Means, Hot-deck Imputation, Last Value Carried Forward, and Predicted Mean imputation.

The Single Imputation option provides a standard range of traditional imputation techniques useful for sensitivity analysis.

Group Means

Imputed values are set to the variable's group mean (or mode in the case of categorical data.)

Hot-deck Imputation

Imputed values are selected from responders that are similar with respect to a set of auxiliary variables.

Last Value Carried Forward (LVCF)

The last observed value of a longitudinal variable is imputed.

Longitudinal variables are those variables intended to be measured at several points in time, such as pre and post test, measurements of an outcome variable made at monthly intervals, laboratory tests made at each visit from baseline, through the treatment period, and through the follow-up period.

For example: if the blood pressures of patients were recorded every month over a period of six months, we would refer to this as one longitudinal variable consisting of six repeated measures or periods.

Linear interpolation is another method for filling in missing values in a longitudinal variable. If a missing value has at least one observed value before, and at least one observed value after, the period for which it is missing, then linear interpolation can be used to fill in the missing value. Although this method logically belongs in the LVCF option, for historical reasons it is only available as an imputation method from within either the Propensity Score Based Method, or the Predictive Model Based Method. For further details see the *Bounded Missing* section.

Predicted Mean

Imputed values are predicted using an ordinary least-squares multiple regression algorithm to impute the most likely value when the variable to be imputed is continuous or ordinal. When the variable to be imputed is a binary or categorical variable, a discriminant method is applied to impute the most likely value.

Multiple Imputation Overview

SOLAS™ 4.0 provides five distinct methods for performing Multiple Imputation:

- The Predictive Model Based Method.
- Propensity Score Based Method.
- Mahalanobis Distance Matching Method.
- Predictive Mean Matching Method
- Propensity Score / Predictive Mean Matching / Mahalanobis Distance Combination Method

Using either method, each missing value is replaced by M ($M \geq 2$) imputed values to create M complete data sets. Multiple Imputation has the following properties:

- ♦ Once the multiple imputations are generated, the resulting data sets can be used by any complete statistical analysis.
- ♦ The extra uncertainty due to missing data is taken into account by imputing two or more different values per missing data entry.

Predictive Model Based Method

The models that are available at present are an Ordinary Least Squares (OLS) Regression, and a Discriminant Model. When the data are continuous or ordinal, the OLS method is applied. When the data are categorical, the discriminant method is applied.

Multiple imputations are generated using a regression model of the imputation variable on a set of user-specified covariates. The imputations are generated via randomly drawn regression model parameters from the Bayesian posterior distribution based on the cases for which the imputation variable is observed.

Each imputed value is the predicted value from these randomly drawn model parameters plus a randomly drawn error-term. The randomly drawn error-term is added to the imputations to prevent over-smoothing of the imputed data. The regression model parameters are drawn from a Bayesian posterior distribution in order to reflect the extra uncertainty due to the fact that the regression parameters can be estimated, but not determined, from the observed data.

Propensity Score Method

The system applies an implicit model approach based on Propensity Scores and an Approximate Bayesian Bootstrap to generate the imputations. The propensity score is the estimated probability that a particular element of data is missing. The missing data are filled in by sampling from the cases that have a similar propensity to be missing. The multiple imputations are independent repetitions from a Posterior Predictive Distribution for the missing data, given the observed data.

Mahalanobis Distance Matching Method

The Mahalanobis distance is used in this method to identify cases that have similar characteristics to cases that have missing values. Missing data are filled in by sampling from the closest cases. The multiple imputations are independent repetitions drawn from the range of closest cases.

Predictive Mean Matching Method

This method applies Ordinary Least Squares Regression for estimating predicted values for each case in the dataset. Rather than using the predicted values for the imputation, they are used to identify similarities between cases with missing values and fully observed cases. Cases are sorted in to Donor Pools and similar to the Propensity Score method imputations are drawn from these pools.

Propensity Score/Predictive Mean Matching/Mahalanobis Distance Combination Method

The Propensity Score method and Predictive Mean Matching method described above are both applied to the data set. This results in each case in the data set having a propensity score and predicted value associated with it. These are then used as covariates and the Mahalanobis Distance method is applied to find cases that can be used to impute missing values.

Single Imputation in SOLAS™ 4.0

Single Imputation is the method in which each missing value in a data set is filled in with one value to yield one complete data set. This allows standard complete-data methods of analysis to be used on the filled-in data set.

Group Means

Missing values in a continuous variable will be replaced with the group mean derived from a grouping variable. The grouping variable must be a categorical variable that has no missing data. Of course if no grouping variable is specified, missing values in the variable to be imputed will be replaced with its overall mean.

When the variable to be imputed is categorical, with different frequencies in two or more categories (providing a unique mode), then the modal value will be used to replace missing values in that variable. Note that if there is no unique mode, *i.e.* if there are equal frequencies in two or more categories, then if the variable is nominal, a value will be randomly selected from the categories with the highest frequency.

If the variable is ordinal, then the 'middle' category will be imputed. If the variable has an even number of categories, a value is randomly chosen from the middle two.

Group Means — Example

This example uses the Fisher (1936) Iris data, **FISHER.MDD**, containing measurements, in centimetres, of sepal length and width, as well as petal length and width, on 50 samples from each of three species of Iris (1=Setosa, 2=Versicolor, 3=Virginica).

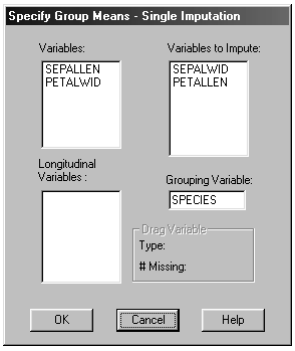
The file **FISHMISS.MDD** is a copy of the original file, created after deleting six values. In this example, we will use Group Means imputation to replace the missing values in the data set.

1. Open the file **FISHMISS.MDD** located in the **SAMPLES** subdirectory.
2. To perform Group Means Imputation, from the datasheet menu bar select **Analyze > Single Imputation**, then choose **Group Means...**

Multiple selection of variables (using drag-and-drop) is supported and is described earlier in this manual.

3. Select the variable(s) you want to impute (**SEPALWID** and **PETALLEN**) by dragging and dropping the variable(s) from the **Variables** list to the **Variables to Impute** field.
4. Drag and drop your grouping variable from the **Variable list** to the **Grouping Variable** field. If you have chosen a grouping variable that has not been previously categorized, the system warns you that you must group the variable. If you do not specify a grouping variable, the overall mean for the variable will be imputed.

For this example, the variables we want to impute are **SEPALWID** and **PETALLEN**, so drag and drop them from the Variables list to the Variables to Impute field. Our grouping variable is the variable **SPECIES**, so this should be dragged to the **Grouping Variable** field.



5. When you are satisfied with your choice, click **OK**. The imputed data set is displayed with the imputed values appearing in pink. This imputed data set can be saved for later analysis, or exported to various other statistics packages (see Chapter 1 – Data Management in the Systems Manual).

Hot-Decking

This method sorts respondents and non-respondents into a number of imputation subsets according to a user-specified set of covariates. An imputation subset comprises cases with the same values as those of the user-specified

covariates. Missing values are then replaced with values taken from matching respondents (*i.e.* respondents that are similar with respect to the covariates). If there is more than one matching respondent for any particular non-respondent, the user has two choices:

1. The first respondent's value as counted from the missing entry downwards within the imputation subset is used to impute. The reason for this is that the first respondent's value may be closer in time to the case that has the missing value. For example, if cases are entered according to the order in which they occur, there may possibly be some type of time effect in some studies.
2. A respondent's value is randomly selected from within the imputation subset. If a matching respondent does not exist in the initial imputation class, the subset will be collapsed by one level starting with the last variable that was selected as a sort variable, until a match can be found. Note that if no matching respondent is found, even after all of the sort variables have been collapsed, three options are available:
   - **Re-specify new sort variables**
     - The user can specify up to five sort variables.
   - **Perform random overall imputation**
     - Where the missing value will be replaced with a value randomly selected from the observed values in that variable.
   - **Do not impute the missing value**
     - SOLAS™ will not impute any missing values for which no matching respondent is found.

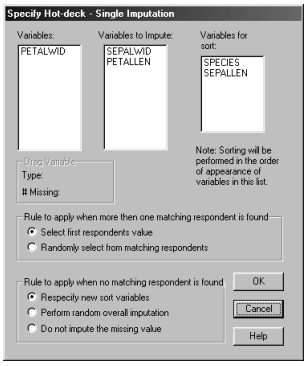
Hot-Decking — Example

This example also uses the data set **FISHMISS.MDD**.

1. Open the file **FISHMISS.MDD**.
2. To perform Hot-deck Imputation, from the datasheet menu bar select **Analyze > Single Imputation**, and **Hot Deck...**
3. Again, the variables we want to impute are **SEPALWID** and **PETALLEN**, so drag them into the **Variables to Impute** field.

For this example we will use **SPECIES** and **SEPALLEN** as our sort variables. The order in which the **Variables for Sort** are specified is important, because if no matching respondent is found in the initial imputation class, the class will be collapsed by one level according to the last variable specified in the **Variables for Sort** field.

4. Since we would expect irises of the same species to be similar with respect to the various measurements, we select **SPECIES** as our primary sort variable, and then select **SEPALLEN** as the secondary sort variable.



5. Under **Rule to apply** when more than one matching respondent is found, choose **Randomly select** from matching respondents.
6. Under **Rule to apply** when no matching respondent is found, choose **Re-specify new sort variables**.
7. When you are satisfied with your choice, click **OK**. The imputed values are displayed in the color orange.

The system sorts the data set in ascending order, so **SPECIES** is sorted first, and **SEPALLEN** is sorted next. Then, for each missing value, the system finds all respondents with matching values for these two variables. Thus, case #96 (which is missing in **SEPALWID**) has **SPECIES=1** and **SEPALLEN=5.0**. There are 7 respondents in this imputation class (cases #1, #42, #55, #56, #68, #73 and #108), with matching values for **SPECIES** and **SEPALLEN**, and a randomly selected respondent is used to impute the missing value.

NOTE: If your sort variables are continuous variables with significant decimal places, exact matches may not occur. You could use the Transform feature to take the integer value of variables that you want to use for sorting.

This imputed data set can be saved for later analysis or exported to any other statistics package. (See Chapter 1 -Data Management in the Systems Manual.)

Last Value Carried Forward

The Last Value Carried Forward (LVCF) technique can be used when the data are longitudinal (*i.e.* repeated measures have been taken per subject). The last observed value is used to fill in missing values at a later point in the study. Therefore one makes the assumption that the response remains constant at the last observed value.

This assumption can be biased if the timing and rate of withdrawal is related to the treatment. For example, in the case of degenerative diseases, using the last observed value to impute for missing data at a later point in the study means that a high observation will be carried forward, resulting in an overestimation of the true end-of-study measurement.

LVCF — Example

This example uses the data set **MI_TRIAL.MDD** (located in the **SAMPLES** subdirectory).

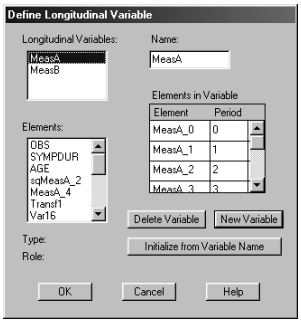
Define Longitudinal Variables

Since LVCF can only be performed on longitudinal variables in SOLAS™, our first step will be to define the Longitudinal Variables in the data set.

1. Open the file **MI_TRIAL.MDD**.
2. From the **Variables** menu, select **Define Variables > Longitudinal...**
3. The system assigns the default name **LVar1** to the first longitudinal variable. Just type **MeasA** into the **Name** field to replace the default name.
4. Click on the variable name in the **Elements** listbox to enable the **Initialize From Variable Name** button, then press this button to include all the "MeasA" variables in the **Elements in Variable** field.
5. The system automatically assigns a period value of zero to the first element, and the remaining elements will be assigned period values of 1, 2, etc. You can change these values by typing in new values.

For example, you might want to change the default period values if your repeated measurement were taken at baseline month1, month6, and month8, i.e. at unequal time intervals. By setting the period values to 1, 6, and 8, you will ensure that linear interpolation of bounded missings will be correct. Here the measurements were taken at month1, month2, and month3, so the default values do not need be changed.

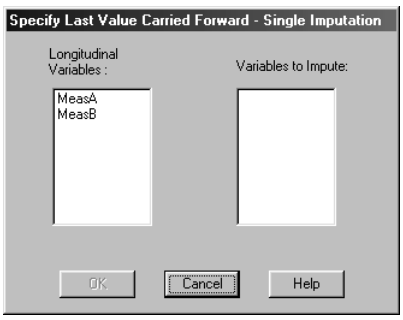
6. Click on **New Variable** to define the elements of our second longitudinal variable.
7. A dialog box appears, asking if you want to save your changes to the longitudinal variable **MeasA**. Click **Yes**.
8. Type the name **MeasB** in the name field, then click on the variable name in the **Elements** listbox to enable the **Initialize From Variable Name** button, then press this button to include all the "MeasB" variables in the **Elements in Variable** field.
9. When you are satisfied that you have defined your longitudinal variable correctly, click **OK** to finish.



LVCF Imputation

1. To perform LVCF Imputation, choose **Single Imputation > Last Value Carried Forward** from the datasheet **Analyze** menu.

2. The two longitudinal variables that we created appear in the Longitudinal Variables list. Drag and drop the variables **MeasA** and **MeasB** from the **Longitudinal Variables** list into the **Variables to Impute** field.



3. When you are satisfied with your choice, click **OK**. The imputed data set is displayed with the imputed values appearing in Blue/Grey.

The value from the last observed period is carried forward to fill in for missing values in later periods. For example; case #7 has a baseline value of 147 for MeasA, but is missing for all subsequent periods. This value of 147 is carried forward to fill in for these missing periods. This imputed data set can be saved for analysis later, or exported to any other statistics package (see *Chapter 1 – Data Management* in the Systems Manual).

Predicted Mean Imputation

Predicted Mean Imputation is performed using an Ordinary Least-Squares Regression or a Discriminant analysis. A general description of these methods is given below.

Ordinary Least-Squares

Using the Least-squares method, missing values are imputed using predicted values from the corresponding covariates using the estimated linear regression models. This method is used to impute all the continuous variables in a data set.

Discriminant

Discriminant Multiple Imputation is a model based method for binary or categorical variables. For each missing data entry, the category with the largest conditional probability given the values of the selected covariates is imputed. More detailed information can be found in Appendix D – Discriminant Multiple Imputation.

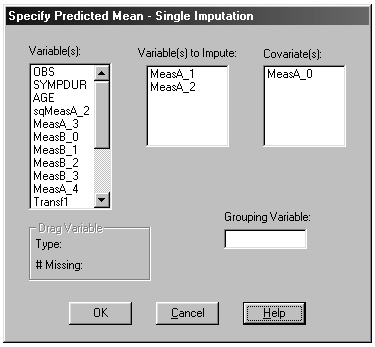
Predicted Mean Imputation – Example

This example uses the data set **MI_TRIAL.MDD** (located in the **SAMPLES** subdirectory).

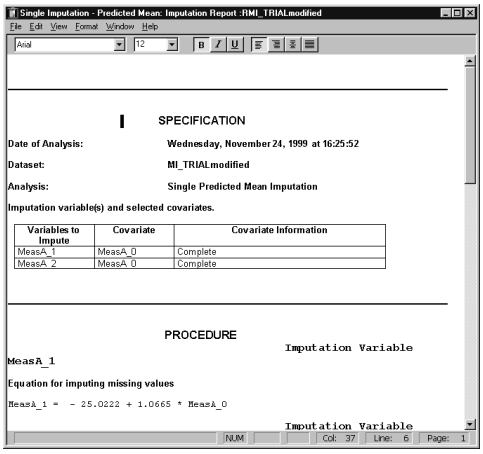
1. Open the datasheet **MI_TRIAL.MDD**, select **Analyze > Single Imputation**, and the **Predicted Mean** option to display the Specify Predicted Mean window.
2. Drag the variables to be imputed, the chosen Covariates, and the Grouping Variable between the **Variable(s)**, the **Variable(s) to Impute**, and the **Covariate(s)** listboxes, and **Grouping Variable** datafield.
3. For this example we have chosen the Variables to be imputed as **MeasA_1** and M**easA_2**, the variable **MeasA_0** as the Covariate.

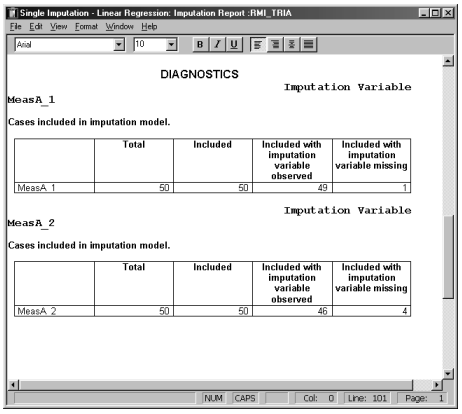
NOTE: You cannot drag variables that do not contain missing values into the "**Variable(s) to Impute**" listbox.

4. When the required variables have been selected, press the **OK** button to display the Specify Predicted Mean window shown below:

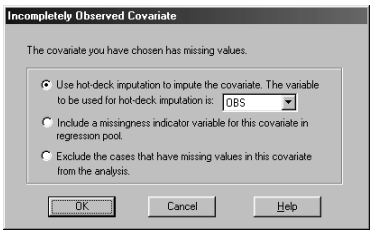


5. After pressing the **OK** button, new datasheet window is displayed where the imputed values are displayed as green text, and an **Imputation Report** (shown in part below) can be selected from the **View** menu:





NOTE: There are no missing values in the variable chosen as the Covariate in this example, but if there were, the following window would be displayed:



Then:

1. If the **Use hot-deck imputation** option is chosen, you must select a variable in the dropdown listbox that will be used to impute the missing values in the Covariate. The dropdown list contains a list of all of the variables in the data set, in the same order as they appear in the datasheet. If more than one matching respondent is found, a value is randomly selected from within the imputation class. If no matching respondent is found, the respondent is selected at random from all the used cases.

2. If the **Include a missingness indicator** is chosen for a covariate *x*, then the independent variable *x* is changed into R_x*x and the intercept is adjusted by adding the independent variable 1-R_x to the regression model, where *Rx* is the response indicator vector for the incomplete covariate *x*. See *Appendix C – Multiple Imputation – Predictive Model Based Method.*

3. If the **Exclude** option is chosen, all of those cases that are missing in the Covariate are excluded, and no missing values will be imputed for these cases.

NOTE: Unless another Covariate is chosen, the Covariate with missing values discussed above will be used in all subsequent steps of the imputation.

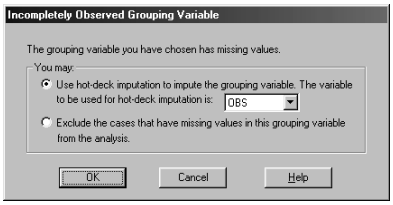
And:

4. If a nominal variable(s) is chosen as the Covariate(s) you will be prompted to create design variables and these will be used in the regression analysis.

5. If there are no groups in the variable chosen as a grouping variable, you will be prompted to group the variable.

NOTE: There are no missing values in the variable chosen as a grouping variable for this example, but if there were,

the following window would be displayed:



Then:

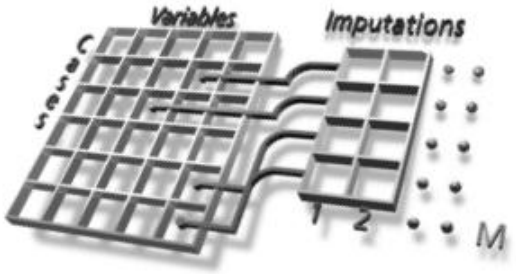
6. If the **Use the hot-deck** option is chosen, you must select a variable in the dropdown listbox that will be used to impute the missing values in the grouping variable. The dropdown list will contain a list of all of the variables in the data set, in the same order as they appear in the datasheet.

If more than one matching respondent is found, a value is randomly selected from within the imputation class. If no matching respondent is found, the respondent is selected at random from all of the used cases.

7. If the **Exclude** option is chosen, all of those cases that are missing in the grouping variable are excluded, and no missing values will be imputed in these cases

Multiple Imputation in SOLAS™ 4.0

Multiple Imputation replaces each missing value in the data set with several imputed values instead of just one. First proposed by Rubin in the early 1970's as a possible solution to the problem of survey non-response, the method corrects the major problems associated with single imputation (see Appendix F, references [1] to [5]). Multiple Imputation creates M imputations for each missing value, thereby reflecting the uncertainty about which value to impute.



The first set of the M imputed values is used to form the first imputed data set, the second set of the M imputed values is used to form the second imputed data set, and so on.

In this way M imputed data sets are obtained. Each of the M imputed data sets is statistically analyzed by the complete-data method of choice. This yields M intermediate results. These M intermediate results are then combined into a final result, from which the conclusions are drawn, according to explicit formulae (see Appendix A). The extra inferential uncertainty due to missing data can be assessed by examining the between imputation variance and the following related measures:

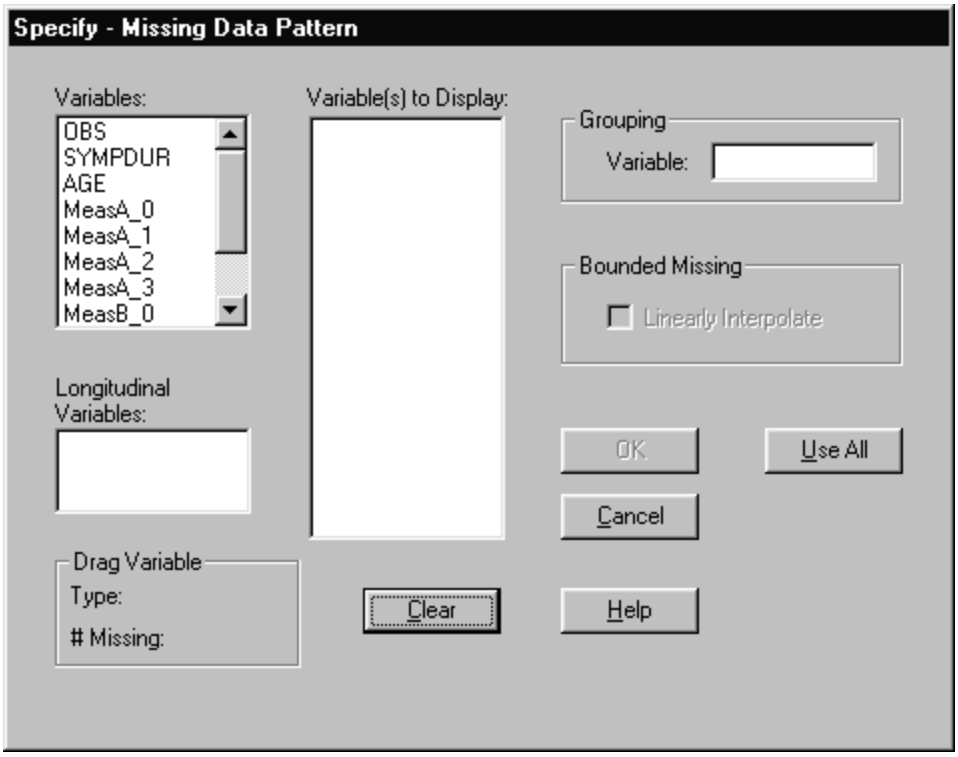
- The relative increases in variance due to non-response (R_m) and the fraction of information missing due to missing data (γ_m).

General

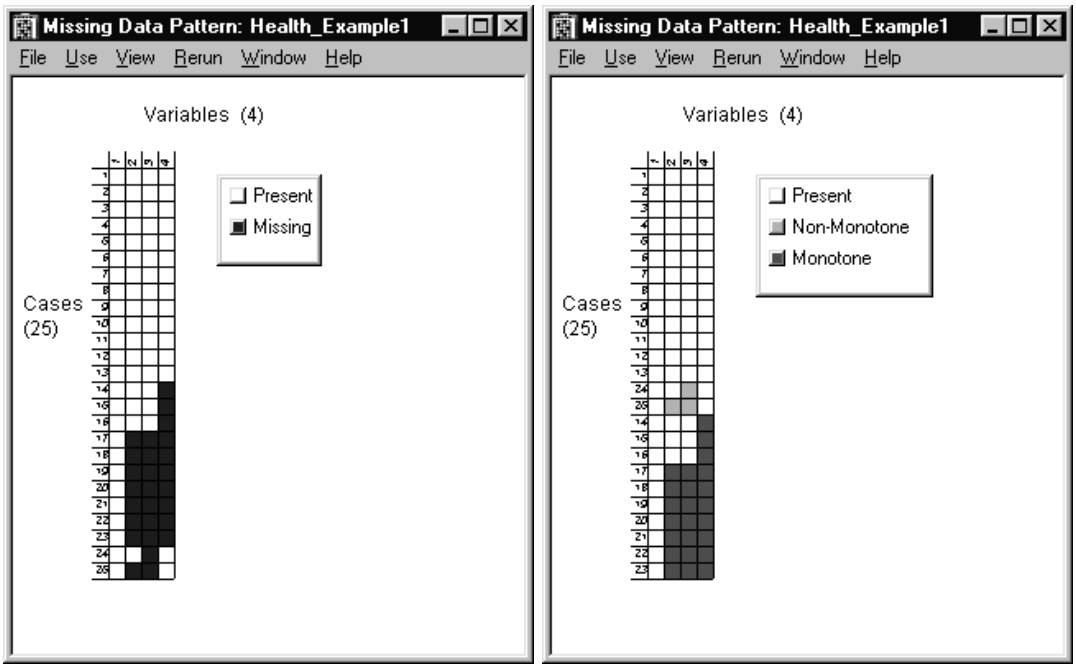
Before the imputations are actually generated, the missing data pattern is sorted as close as possible to a Monotone missing data pattern, and each missing data entry is labelled as either Monotone missing or Non-monotone missing, according to where it fits in the sorted missing data pattern.

Missing Data Pattern

The Missing Data Pattern window displays missing data patterns from your data set before and after imputation. You can display the **Specify Missing Data Pattern** window (shown below) from the **View** menu of a datasheet. Using this window you specify which variables should be used to determine a missing data pattern. You can also specify a grouping variable, in which case separate patterns will be generated for each group.

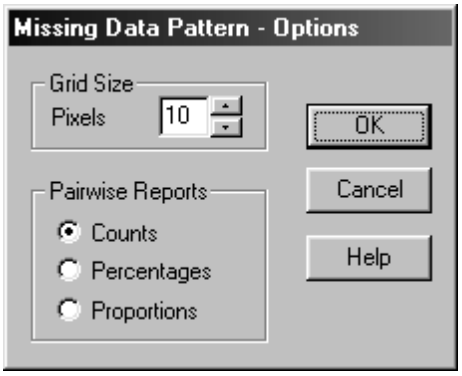


After specifying the variables to use and pressing the **OK** button, a **Missing Data Pattern** window is displayed (below left) with the missing pattern before imputation. From the **View** menu of a **Missing Data Pattern** window you can display the **Monotone** pattern (below right). You can also display a legend from which you can easily identify the missing data type(s).



From the **View** menu of a **Missing Data Pattern** window you can display **Pairwise Missingness/Presence**. These display a matrix that contains the number of cases that are missing/present in each pair of variables.

If you right-click on any of the cells in the missing pattern, a new panel will display the case number, the variable name, and its status. Also from the **View** menu, you display an **Options** window which allows you to choose between various options to use in the display.



The third view of your data set is displayed from the **View** menu of the Output pages after you have performed the imputation (see *Multiple Imputation - Output* later in this manual). You can also use the **View** menu **Legend** option to display a colored legend that identifies the method of imputation used for the missing data.

Monotone Missing Data Pattern

A monotone missing data pattern occurs when the variables can be ordered, from left to right, such that a variable to the left is at least as observed as all variables to the right. For example, if variable A is fully observed and variable B is sometimes missing, A and B form a monotone pattern. Or if A is only missing if B is also missing, A and B form a monotone pattern. If A is sometimes missing when B is observed, and B is sometimes missing when A is observed, then the pattern is not monotone (e.g. see Little and Rubin, 1987, Section 6.4, and References [6] and [7] in *Appendix F*.)

We also distinguish between a missing data pattern and a local missing data pattern:

- A missing data pattern refers to the entire data set, such as a Monotone missing data pattern.
- A local missing data pattern for a case refers to the missingness for a particular case of a data set.
- A local missing data pattern for a variable refers to the missingness for that variable.

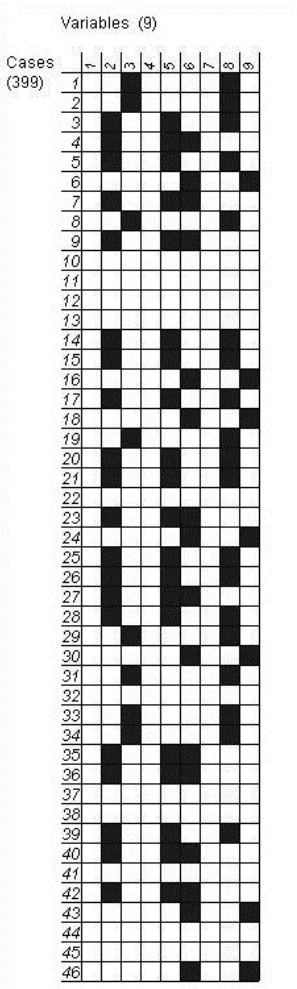
NOTE: If two cases have the same sets of observed variables and the same sets of missing variables, then these two cases have the same local missing data pattern.

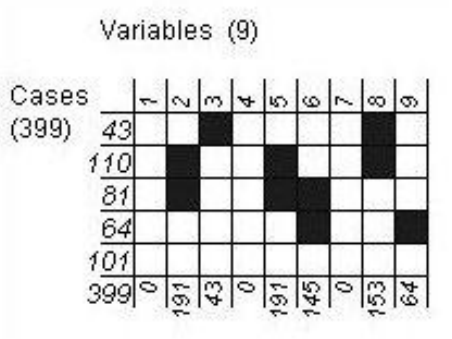
A Monotone pattern of missingness, or a close approximation to it can be quite common. For example, in longitudinal studies, subjects often drop out as the study progresses so that all subjects have time 1 measurements, a subset of subjects have time 2 measurements, only a subset of those subjects have time 3 measurements, and so on.

SOLAS™ sorts variables and cases into a pattern that is as close as possible to a Monotone pattern. Monotone patterns are attractive, because the resulting analysis is flexible. The resulting imputation is completely principled since only observed/real data are used in the model to generate the imputed values. See Rubin, 1987, Chapter 5.

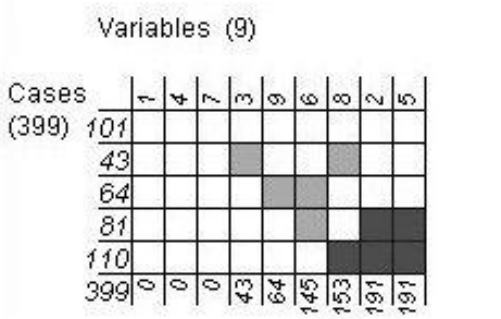
Collapse Missing Data Pattern

For larger data sets it can be difficult to interpret the missing data pattern. In SOLAS™ 4.0 there is a new option to collapse a missing data pattern. This option looks at the pattern of missing and observed variables for each case. It displays the various patterns that occur in the data indicating how many times that particular pattern occurs.





Collapsed missing data pattern



Collapsed monotone missing data pattern

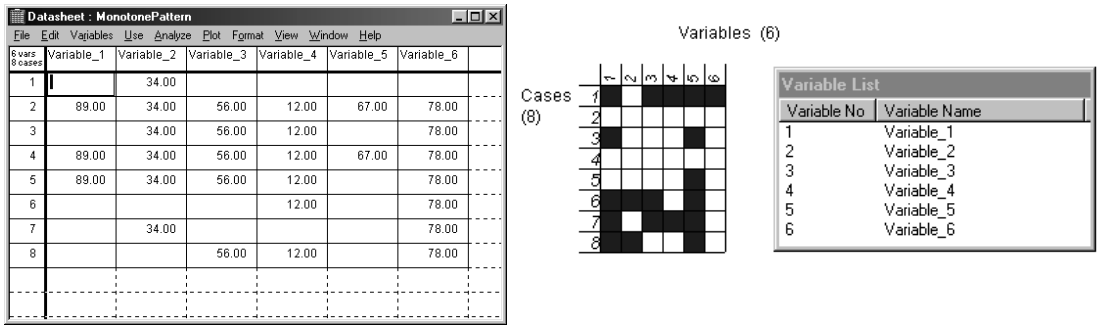
The left hand image above shows the first 46 cases in the standard missing data pattern. There are a total of 399 cases in the dataset so the user has to scroll down the screen to see the entire pattern. However it is now possible to collapse the pattern to make it more readable. From the **View** menu you can select **Collapse** and this will condense the pattern into the minimal display possible. From the image on the left it can be seen that cases 1, 2, 8, 19, 29, 31, 33 and 34 all contain missing values for variables #3 and #8. From the collapsed image on the right we see that this pattern occurs 43 times throughout the entire dataset.

In the collapsed image the first column indicates the number of rows that have that particular pattern. The final row gives the total number of missing values for that variable.

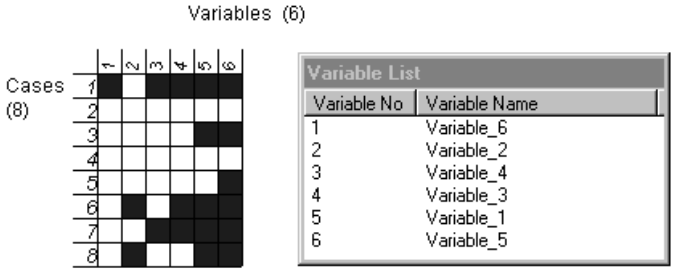
Example of Sorting into a Monotone Missing Data Pattern

In SOLAS™ 4.0, finding a Monotone missing data pattern consists of three main processes. The first process sorts the variables in a datasheet from the most observed to the least observed.

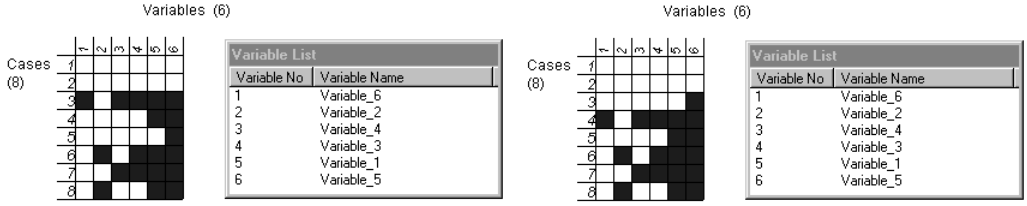
This is demonstrated using a simple datasheet and the Variables List window. By selecting the **View** menu in a datasheet window, you can display the Missing Data Pattern, and from the **View** menu in the Missing Data Pattern window you can display the Variable List windows as shown in this example. The datasheet, the unsorted data in the Missing Data Pattern window and the unsorted Variable List window are shown below:



The new pattern (after sorting) can be viewed in the Missing Pattern window and in the Variable list window. Missing cases are represented by the darkened squares. All variables with the same local missing data pattern are adjacent. After sorting the variables from most observed to least observed in the first process, we have the following result.

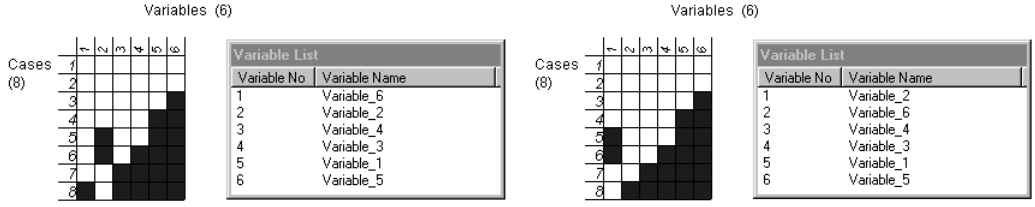


The second process rearranges the cases. Starting with the least missing variable (#6), cases with the most missing values are moved towards the bottom of the sort order, and this process is repeated for the next least missing variable (#5) as shown in the left and right-hand images below:



The same process is continued for variables #4, #3, and #2 as shown in the left-hand image below. All cases with the same local missing data pattern are adjacent.

Finally, an additional scan is performed to determine whether any of the variables that lie outside the Monotone pattern can be moved in order to include more missing values in the Monotone pattern. In this example, swapping the first two variables results in extra missing values being included in the Monotone pattern. The result of this process is shown in the right-hand image below.



The right-hand image above displays the final result in constructing an approximate Monotone pattern for the example datasheet shown earlier. The missing values in the lower right corner are labelled as Monotone missing, and the others as Non-monotone missing.

Predictive Model Based Method

If Predictive Model Based Multiple Imputation is selected, then an ordinary least-squares regression method of imputation is applied to the continuous, integer, and ordinal imputation variables, and discriminant multiple imputation is applied to the nominal imputation variables.

Ordinary Least-squares Regression

The predictive information in a user-specified set of covariates is used to impute the missing values in the variables to be imputed. First, the Predictive Model is estimated from the observed data. Using this estimated model, new linear regression parameters are randomly drawn from their Bayesian posterior distribution. The randomly drawn values are used to generate the imputations, which include random deviations from the model's predictions.

Drawing the exact model from its posterior distribution ensures that the extra uncertainty about the unknown true model is reflected.

In the system, multiple regression estimates of parameters are obtained using the method of least squares. If you have declared a variable to be nominal, then you need design variables (or dummy variables) to use this variable as a predictor variable in a multiple linear regression. The system's multiple regression allows for this possibility and will create design variables for you.

Generation of Imputations

Let Y be the variable to be imputed, and let X be the set of covariates. Let Y_{obs} be the observed values in Y, and Y_{mis} the missing values in Y. Let X_{obs} be the units corresponding to Y_{obs}

The analysis is performed in two steps:

1. The Linear Regression Based Method regresses Y_{obs} on X_{obs} to obtain a prediction equation of the form: $Y_{mis} = a + bX_{mis}$.
2. A random element is then incorporated in the estimate of the missing values for each imputed data set. The computation of the random element is based on a posterior drawing of the regression coefficients and their residual variances.

Refer to *Appendix C* for more detailed information about the analysis that is performed for Multiple Imputation using the Predictive Model Based Method.

Posterior Drawing of Regression Coefficients and Residual Variance

Parameter values for the regression model are drawn from their posterior distribution given the observed data using non-informative priors. In this way, the extra uncertainty due to the fact that the regression parameters can be estimated, but not determined, from Y_{obs} and X_{obs} is reflected.

Using estimated regression parameters rather than those drawn from its posterior distribution can produce improper results, in the sense that the between imputation variance is underestimated.

For more detailed information see *Appendix C – Multiple Imputation – Predictive Model Based Method.*

Discriminant Multiple Imputation

Discriminant multiple imputation is a model based method for imputing binary or categorical variables.

Let i,…,s be the categories of the categorical imputation variable y. Bayes' Theorem is used to calculate the probability that a missing value in the imputation variable y is equal to its j^{th} category given the set of the observed values of the covariates and of y.

For more details see *Appendix D – Discriminant Multiple Imputation*.

Propensity Score Method

The system applies an implicit model approach based on Propensity Scores and an Approximate Bayesian Bootstrap to generate the imputations. The underlying assumption about Propensity Score Multiple Imputation is that the non-response of an imputation variable can be explained by a set of covariates using a logistic regression model. The multiple imputations are independent repetitions from a Posterior Predictive Distribution for the missing data, given the observed data.

Variables are imputed from left to right through the data set, so that values that are imputed for one variable can be used in the prediction model for missing values occurring in variables to the right of it. The system creates a temporary variable that will be used as the dependent variable in a logistic regression model. This temporary variable is a response indicator and will equal 0 for every case in the imputation variable that is missing and will equal 1 otherwise.

The independent variables for the model will be a set of baseline/fixed covariates that we think are related to the variable we are imputing. For example, if the variable being imputed is period t of a longitudinal variable, the covariates might include the previous periods (t-1, t-2,...).

The regression model will allow us to model the "missingness" using the observed data. Using the regression coefficients, we calculate the propensity that a subject would have a missing value in the variable in question. In other words, the propensity score is the conditional probability of "missingness", given the vector of observed covariates.

Each missing data entry of the imputation variable *y* is imputed by values randomly drawn from a subset of observed values of *y*, i.e. its donor pool, with an assigned probability close to the missing data entry that is to be imputed. The Donor Pool defines a set of cases with observed values for that imputation variable.

Defining Donor Pools Based on Propensity Scores

Using the options in the Donor Pool window, the cases of the data sets can be partitioned into *c* donor pools of respondents according to the assigned propensity scores, where *c*=5 is the default value of *c*. This is done by sorting the cases of the data sets according to their assigned propensity scores in ascending order.

The Donor Pool page gives the user more control over the random draw step in the analysis.
You are able to set the sub-set ranges and refine these ranges further using another variable known as the Refinement Variable that is described below.

Three ways of defining the Donor Pool sub-classes are provided:

1. You can divide the sample into c equal sized subsets; the default will be 5. If the value of c results in not more than 1 case being available to the selection algorithm, c will decrement by 1 until such time as there is sufficient data. The final value of c used is included in the Imputation Report output described later in this manual.
2. You can use the subset of c cases that are closest with respect to propensity score. This option allows you to specify the number of cases that are to be included in the sub-class. The default c will be 10 and cannot be set to a value less than 2. If less than 2 cases are available, a value of 5 will be used for c.
3. You can use the subset of d% of the cases that are closest with respect to propensity score. This is the percentage of "closest" cases in the data set to be included in the sub-class. The default for d will be 10.00 and cannot be set to a value that will result in less than 2 cases being available. If less than 2 cases are available, a d value of 5 will be used.

Refer to *Appendix E – Propensity Score Multiple Imputation* for more detailed information.

Mahalanobis Distance Matching Method

The Mahalanobis distance is a metric that can be used to measure the dissimilarity between two vectors. In this case, the vectors will be cases from the dataset and they will be composed of the values from the covariates specified for the calculation.

Generation of Imputations

Consider that $\vec{y}$ represents the vector for the case containing the missing value and $\vec{x}_i$ is a complete case. The distance between these is calculated as follows:

$$d(\vec{x}_i, \vec{y}) = \sqrt{(\vec{x}_i - \vec{y})^T S^{-1} (\vec{x}_i - \vec{y})}$$

where S is the covariance matrix.

Each missing value from the imputation variable *y* is imputed by values randomly drawn from a subset of observed values, i.e. its donor pool, with the shortest Mahalanobis distance to the missing data entry that is to be imputed. The Donor Pool defines a set of cases with observed values for that imputation variable.

Defining Donor Pools Based on Mahalanobis Distances

The Donor Pool page gives the user control over the random draw step in the analysis. You are able to set the sub-set ranges and refine these ranges further using another variable known as the Refinement Variable that is described below.

Two ways of defining the Donor Pool sub-classes are provided:

1. You can use the subset of c cases that are closest with respect to Mahalanobis distance. This option allows you to specify the number of cases that are to be included in the sub-class. The default c will be 10 and cannot be set to a value less than 2. If less than 2 cases are available, a value of 5 will be used for c.
2. You can use the subset of d% of the cases that are closest with respect to Mahalanobis distance. This is the percentage of "closest" cases in the data set to be included in the sub-class. The default for d will be 10.00 and cannot be set to a value that will result in less than 2 cases being available. If less than 2 cases are available, a d value of 5 will be used.

Predictive Mean Matching Method

If Predictive Mean Matching Multiple Imputation is selected, then an ordinary least-squares regression method is applied to the continuous, integer, and ordinal imputation variables, and discriminant multiple imputation is applied to the nominal imputation variables.

The predictive information in a user-specified set of covariates is used to impute the missing values in the variables to be imputed. First, the Predictive Model is estimated from the observed data. There is an option to use either the estimated model or using this estimated model, draw new linear regression parameters randomly from their Bayesian posterior distribution. The randomly drawn values are used to generate the imputations, which include random deviations from the model's predictions. Drawing the exact model from its posterior distribution ensures that the extra uncertainty about the unknown true model is reflected.

In the system, multiple regression estimates of parameters are obtained using the method of least squares. If you have declared a variable to be nominal, then you need design variables (or dummy variables) to use this variable as a predictor variable in a multiple linear regression. The system's multiple regression allows for this possibility and will create design variables for you.

Generation of Imputations

Let Y be the variable to be imputed, and let X be the set of covariates. Let Y_{obs} be the observed values in Y, and Y_{mis} the missing values in Y. Let X_{obs} be the units corresponding to Y_{obs}. The Linear Regression Based Method regresses Y_{obs} on X_{obs} to obtain a prediction equation of the form: $Y = a + bX$. Predicted values are then estimated for all cases in the dataset, regardless of whether they have values missing or not. These predictions are then used to create donor pools.

Defining Donor Pools Based on Predicted Values

Using the options in the Donor Pool window, the cases of the data sets can be partitioned into *c* donor pools of respondents according to the assigned predicted values, where c=5 is the default value of *c*. This is done by sorting the cases of the data sets according to their assigned predicted values in ascending order.

The Donor Pool page gives the user more control over the random draw step in the analysis. You are able to set the sub-set ranges and refine these ranges further using another variable known as the Refinement Variable that is described below.

Three ways of defining the Donor Pool sub-classes are provided:

1. You can divide the sample into c equal sized subsets; the default will be 5. If the value of c results in not more than 1 case being available to the selection algorithm, c will decrement by 1 until such time as there is sufficient data. The final value of c used is included in the Imputation Report output described later in this manual.
2. You can use the subset of c cases that are closest with respect to propensity score. This option allows you to specify the number of cases that are to be included in the sub-class. The default c will be 10 and cannot be set to a value less than 2. If less than 2 cases are available, a value of 5 will be used for c.
3. You can use the subset of d% of the cases that are closest with respect to predicted value. This is the percentage of "closest" cases in the data set to be included in the sub-class. The default for d will be 10.00 and cannot be set to a value that will result in less than 2 cases being available. If less than 2 cases are available, a d value of 5 will be used.

Propensity Score/Predictive Mean/Mahalanobis Distance Combination Method

The system employs the three methods outlined previously to generate imputations. Propensity Scores (p. 22) and Predicted Values (p.23) are calculated for all cases in the dataset. The same set of covariates is used for both calculations. Once these calculations are completed the propensity scores and predicted values are then treated as additional variables in the dataset and they are used as the covariates for the Mahalanobis Distance method (p. 22).

Refinement variable

For all of the methods that use a donor pool to generate imputations it is possible to specify a refinement variable. Using the Donor Pool window, a refinement variable *w* can be chosen, and can be applied to each of the Donor Pool options described above. For each missing value of *y* that is to be imputed, a smaller sub-set is selected on the basis of the association between *y* and *w.* This smaller sub-set will then be used to generate the imputations.

For each missing value of *y*, the imputations are randomly drawn according to the Approximate Bayesian Bootstrap method from the chosen sub-set of observed values of *y*.

Using this method (also described in Rubin, (1987) Multiple Imputation for Nonresponse in Surveys, referenced in *Appendix F* [1]), a random sample (with replacement) is randomly drawn from the chosen sub-set of observed values to be equal in size to the number of observed values in this sub-set. The imputations are then randomly drawn from this sample.

The Approximate Bayesian Bootstrap method is applied in order to reflect the extra uncertainty about the predictive distribution of the missing value of *y*, given the chosen sub-set of observed values of y. This predictive distribution can be estimated from the chosen sub-set of observed values of *y*, but not determined. Drawing the imputations randomly from the chosen sub-set of observed values rather than applying the Approximate Bayesian Bootstrap, would result in improper imputation in the sense that the between imputation variance is underestimated.

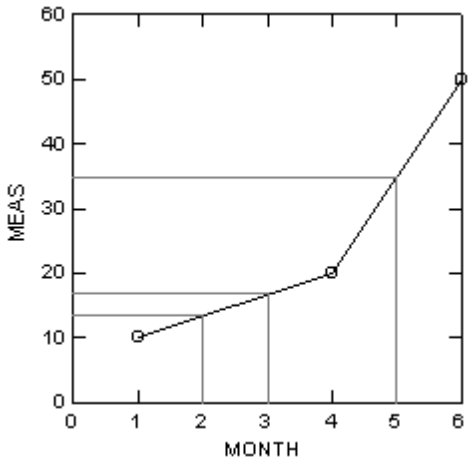
Bounded Missing

This type of missing value can only occur when a variable is longitudinal. It is a missing value that has at least one observed value before, and at least one observed value after the period for which it is missing. The following table shows an example of bounded missing values. The variables Month1 to Month6 are a set of longitudinal measures.

Patient	Month1	Month2	Month3	Month4	Month5	Month6
101	10	*	*	20	*	50
102	20	40	*	30	*	*
103	30	*	*	*	*	50

* = missing, shaded = bounded missing

Linear interpolation can be used to fill in missing values that are longitudinal variables. So for example, using linear interpolation, patient 101's missing values for months 2, 3, and 5 would be imputed as follows:



So the imputed value for month 2 will be 13.33, the imputed value for month 3 will be 16.67, and for month 5 will be 35.

Generating the Multiple Imputations

After the missing data pattern is sorted and the missing data entries are either labelled as Non-monotone missing or Monotone missing, the imputations are generated in two steps.

1. The Non-monotone missing data entries are imputed first.
2. Then the Monotone missing data entries are imputed using the previously imputed data for the Non-monotone missing data entries.

The Non-monotone missing data entries are always imputed using a Predictive Model Based Multiple Imputation. The Monotone missing data entries are imputed by the user-specified method, which can be either the Predictive Model Based method, the Propensity Score method, the Mahalanobis Distance Matching method, the Predictive Mean Matching method or the Combination method.

Covariates that are used for the generation of the imputations are selected for each imputation variable separately. For each imputation variable, two sets of covariates are selected. One set of covariates is used for imputing the Non-monotone missing data entries and the other set of covariates is used for imputing the Monotone missing data entries in that variable. After the missing data pattern is sorted, the missing data entries are labelled as Non-monotone or Monotone. For both sets of selected covariates for an imputation variable, a special subset is the fixed covariates.

Fixed covariates are all selected covariates other than imputation variables and are used for the imputation of missing data entries for Monotone and Non-monotone missing patterns. This is only the case for fixed covariates.

Imputing the Non-monotone Missing Data

The Non-monotone missing data are imputed for each sub-set of missing data by a series of individual linear regression multiple imputations (or discriminant multiple imputations) using as much as possible observed and previously imputed data. Information about Linear Regression and Discriminant Multiple Imputation in SOLAS™ 4.0 can be found *in Appendix C – Multiple Imputation – Predictive Model Based Method*.

First, the leftmost Non-monotone missing data are imputed. Then the second leftmost Non-monotone missing data are imputed using the previously imputed values. This process continues until the rightmost Non-monotone missing data are imputed using the previously imputed values for the other Non-monotone missing data in the same sub-set of cases.

The user can specify, or add, covariates for use in the Predictive Model for any variables that will be imputed. More information about using covariates is given in the example below.

Imputing the Monotone Missing Data

The Monotone missing data are sequentially imputed for each set of imputation variables with the same local pattern of missing data. First the leftmost set is imputed using the observed values of this set and its selected fixed covariates only. Then the next set is imputed using the observed values of this set, the observed and previous imputed values of the first set, and the selected fixed covariates.

This continues until the Monotone missing data of the last set is imputed. For each set, the observed values of this set, the observed and imputed values of the previously imputed sets, and the fixed covariates are used. If multivariate propensity score multiple imputation is selected for the imputation of the Monotone missing data, then this method is applied for each subset of sets having the same local missing data pattern.

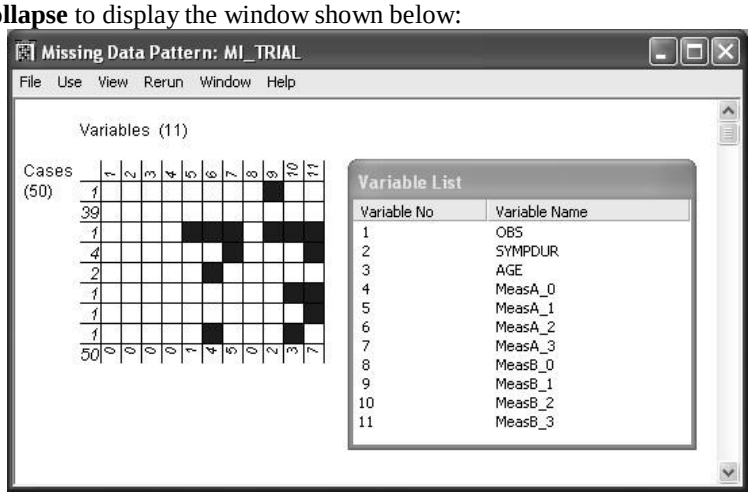
Short Examples

These short examples use the data set **MI_TRIAL.MDD** (located in the **SAMPLES** subdirectory). This data set contains the following 11 variables measured for 50 patients in a clinical trial:

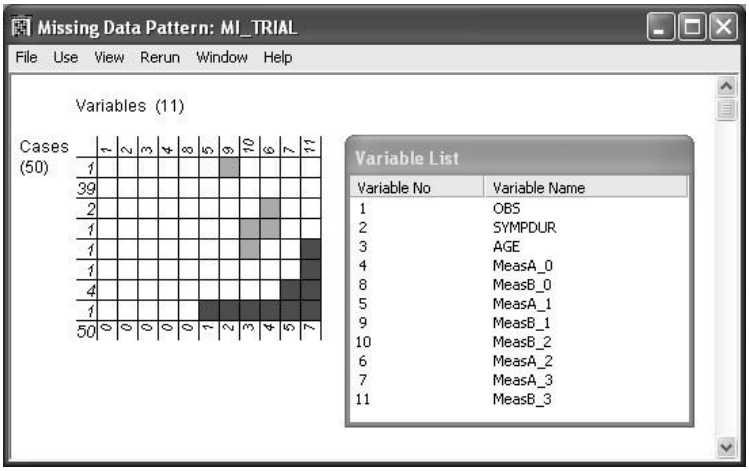
- **OBS** - Observation number.
- **SYMPDUR** - Duration of symptoms.
- **AGE** - The patient's age.
- **MeasA_0**, **MeasA_1**, **MeasA_2**, and **MeasA_3**. The baseline measurement for the response variable MeasA and three post-baseline measurements taken at month 1, month 2, and month 3.
- **MeasB_0**, **MeasB_1**, **MeasB_2**, and **MeasB_3**. The baseline measurement for the response variable MeasB and three post-baseline measurements taken at month 1, month 2, and month 3.

The variables **OBS**, **SYMPDUR**, **AGE**, **MeasA_0**, and **MeasB_0** are all fully observed, and the remaining 6 variables contain missing values. To view the missing pattern for this data set, do the following:

1. From the datasheet window, select **View** and **Missing Data Pattern…**. In the Specify Missing Data Pattern window, press the **Use All** button. From the **View** menu of the Missing Data Pattern window select **Collapse** to display the window shown below:



2. From the **View** menu of the Missing Data Pattern window, select **View Monotone Pattern.**

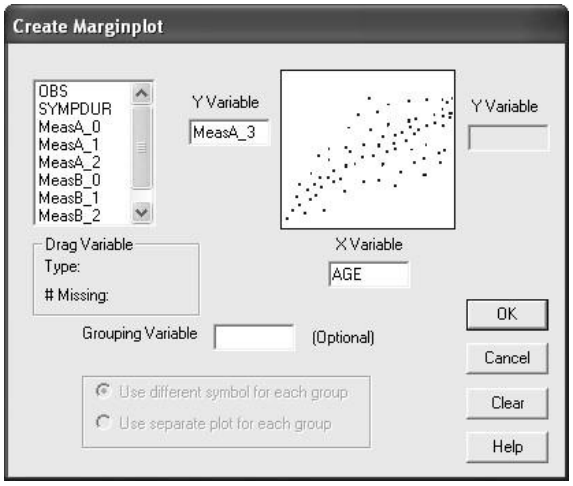


Note that after sorting the data into a Monotone pattern, the time structure of the longitudinal measures is preserved, so the missing data pattern in this data set is Monotone over time.

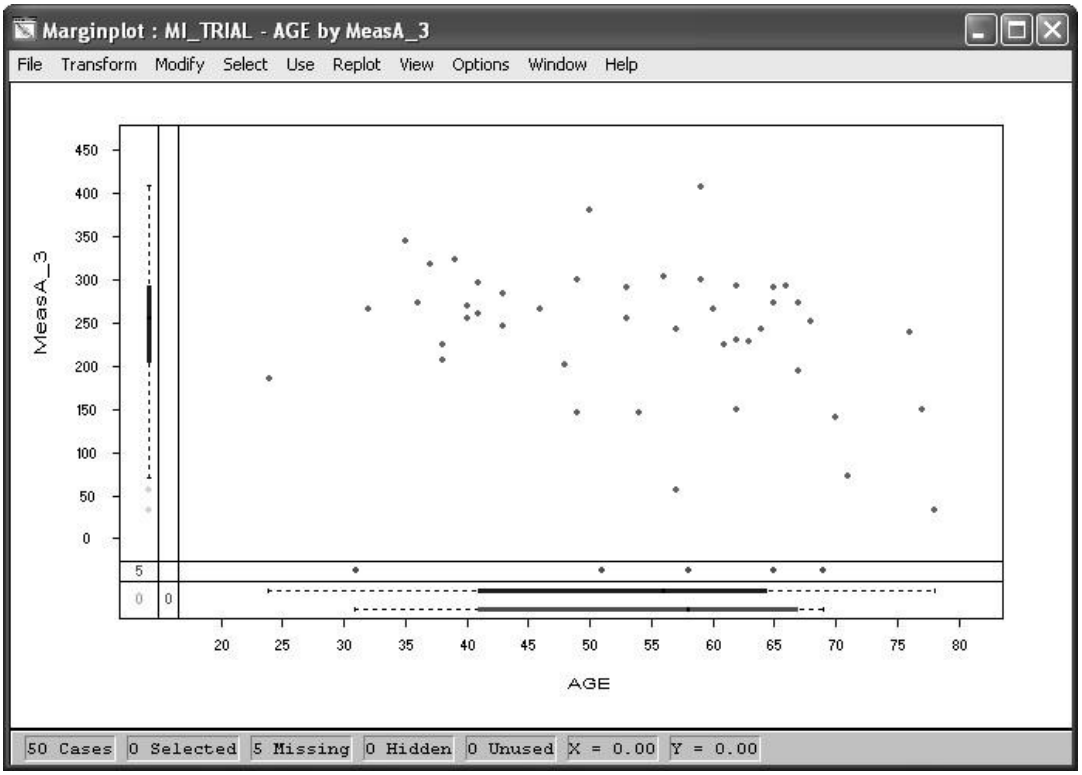
3. To close the **Missing Data Pattern** window, select **File** and **Close**.

It is also possible to plot variables that contain missing variables.

1. From the **Plot** menu select **Marginplot…**



2. Select which variables are to be plotted using the drag and drop method.



3. The fully observed cases are plotted as a normal scatterplot. On the X-axis there are two box-plots. The blue (upper) box represents the observed values. The red (lower) box represents the values that have an observed value for **AGE** but none for **MeasA_3**. The same types of boxplots are available on the Y-axis but because the **AGE** variable is fully observed there is only one boxplot present. This allows the user to see how the cases with missing values are distributed within other variables.

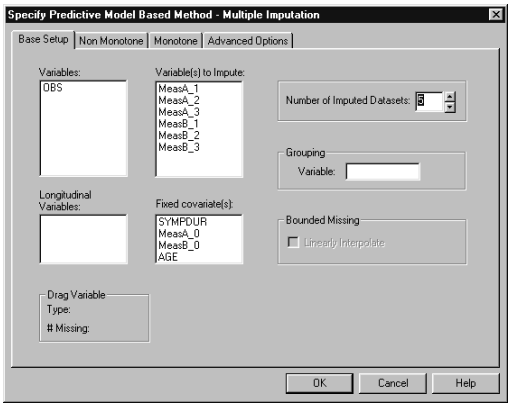
Predictive Model Based Method – Example

We will now multiply impute all of the missing values in this data set using the Predictive Model Based Method by executing the following steps:

1. From the **Analyze** menu, select **Multiple Imputation** and **Predictive Model Based Method**.
2. The Specify Predictive Model window is displayed. The window opens with two pages or tabs: **Base Setup** and **Advanced Options**. As soon as you select a variable to be imputed, a **Non-Monotone** tab and a **Monotone** tab are also displayed.

Base Setup

Selecting the **Base setup** tab allows you specify which variables you want to impute, and which variables you want to use as covariates for the predictive model.

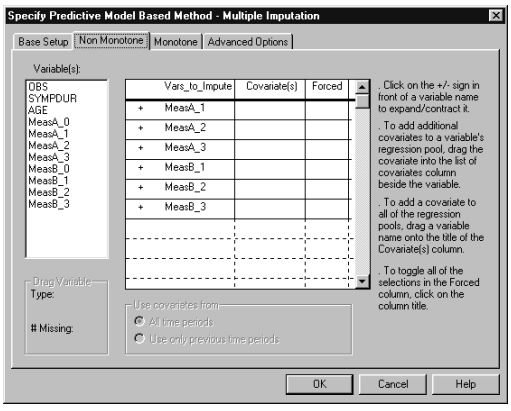


1. Using the datasheet **MI_TRIAL**, drag-and-drop the variables **MeasA_1**, **MeasA_2**, **MeasA_3**, **MeasB_1**, **MeasB_2**, **MeasB_3** into the Variables to Impute field.
2. Drag and drop the variables **SYMPDUR**, **AGE**, **MeasA_0**, and **MeasB_0** into the Fixed Covariates field.
3. As there is no Grouping variable in this data set, we can leave this field blank.

Non-Monotone

Selecting the **Non-monotone** tab allows you to add or remove covariates from the predictive model used for imputing the non-monotone missing values in the data set. (These can be identified in the **Missing Data Pattern** mentioned earlier.)

You select the + or - signs to expand or contract the list of covariates for each imputation variable.



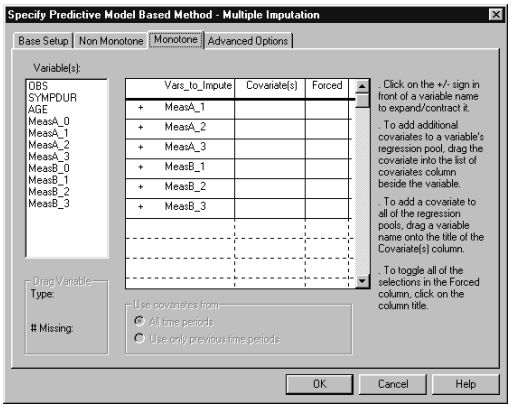
For each imputation variable, the list of covariates will be made up of the variables specified as Fixed Covariates in the **Base Setup** tab, and all of the other imputation variables. Variables can be added and removed from this list of covariates by simply dragging and dropping the variable from the covariate list to the variables field, or vice versa. Even though a variable appears in the list of covariates for a particular imputation variable, it might not be used in the final model.

The program first sorts the variables so that the missing data pattern is as close as possible to monotone, and then, for each missing value in the imputation variable, the program works out which variables, from the total list of covariates, can be used for prediction.

By default, all of the covariates are forced into the model. If you uncheck a covariate, it will not be forced into the model, but will be retained as a possible covariate in the stepwise selection. Details of the models that were actually used to impute the missing values are included in the **Output Log** that can be selected from the **View** menu of the Multiply-Imputed Data Pages. These data pages will be displayed after you have specified the imputation and pressed the **OK** button in the Specify Predictive Model window.

Monotone

Selecting the **Monotone** tab allows you to add or remove covariates from the predictive model used for imputing the monotone missing values in the data set. (These can be identified in the **Missing Data Pattern** mentioned earlier.)

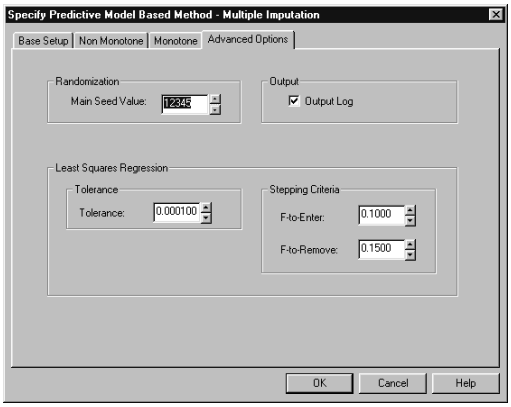


Again, you select the + or - signs to expand or contract the list of covariates for each imputation variable.

For each imputation variable, the list of covariates will be made up of the variables specified as Fixed Covariates in the **Base Setup** tab, and all of the other imputation variables. Variables can be added and removed from this list by simply dragging and dropping. Even though a variable appears in the list of covariates for a particular imputation variable, it might not be used in the final model. The program first sorts the variables so that the missing data pattern is as close as possible to monotone, and then uses only the variables that are to the left of the imputation variable as covariates. Details of the models that were actually used to impute the missing values are included in the **Output Log**.

Advanced Options

Selecting the **Advanced Options** tab displays a window that allows you to choose control settings for the regression/discriminant model.



Randomization

Main Seed Value

The **Main Seed Value** is used to perform the random selection within the propensity subsets. The default seed is 12345. If you set this field to blank, or set it to zero, then the clock time is used.

Output Log
The Output Log is a comprehensive list of regression equations etc. that have been calculated for the imputed variable(s).

Least Squares Regression

Tolerance
The value set in the **Tolerance** datafield controls numerical accuracy. The tolerance limit is used for matrix inversion to guard against singularity. No independent variable is used whose R^2 with other independent variables exceeds (1-Tolerance). You can adjust the tolerance using the scrolled datafield.

Stepping Criteria
Here you can select **F-to-Enter** and **F-to-Remove** values from the scrolled datafields, or enter your chosen value. If you wish to see more variables entered in the model, set the **F-to-Enter** value to a smaller value. The numerical value of **F-to-remove** should be chosen to be less than the **F-to-Enter** value.

Output

When you are satisfied that you have specified your analysis correctly, click the **OK** button. The multiply-imputed datapages will be displayed, with the imputed values appearing in Red or Blue. Refer to "Analyzing Multiple Imputed Data sets" (p. 49) for further details of analyzing these data sets and combining the results.

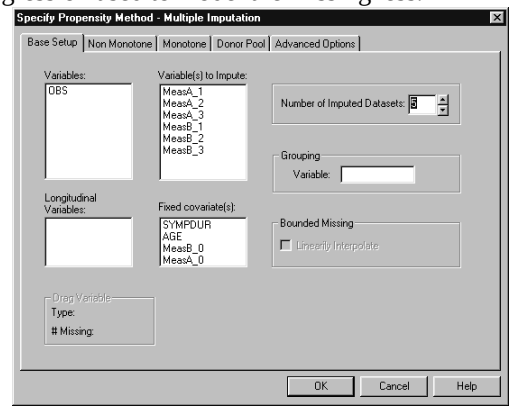
Propensity Score Method — Example

We will now multiply impute all of the missing values in the data set using the Propensity Score Based Method.

1. From the **Analyze** menu, select **Multiple Imputation** and **Propensity Score Method**.
2. The **Specify Propensity Method** window is displayed and is a tabbed (paged) window. The window opens with two pages or tabs: **Base Setup** and **Advanced Options**. As soon as you select a variable to be imputed, a **Non-Monotone** tab, a **Monotone** tab, and a **Donor Pool** tab are also displayed.

Base Setup

Selecting the **Base Setup** tab allows you specify which variables you want to impute, and which variables you want to use as covariates for the logistic regression used to model the missingness.

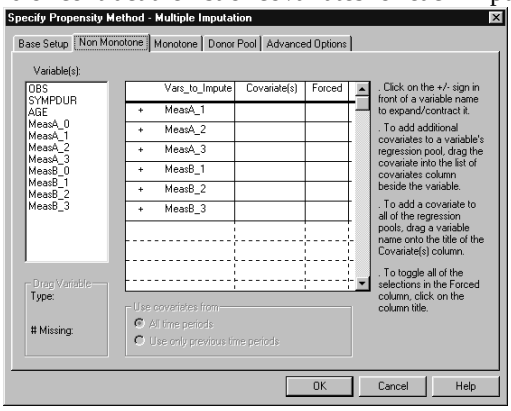


1. Drag-and-drop the variables **MeasA_1**, **MeasA_2**, **MeasA_3**, **MeasB_1**, **MeasB_2**, **MeasB_3** into the Variables to Impute field.
2. Drag and drop the variables **SYMPDUR**, **AGE**, **MeasA_0**, and **MeasB_0** into the Fixed Covariates field.
3. As there is no Grouping variable in this data set, we can leave this field blank.

Non-Monotone

Selecting the **Non-monotone** tab allows you to add or remove covariates from the logistic model used for imputing the non-monotone missing values in the data set. (These can be identified in the **Missing Data Pattern** mentioned earlier in the Predictive Model example.)

You select the + or - signs to expand or contract the list of covariates for each imputation variable.



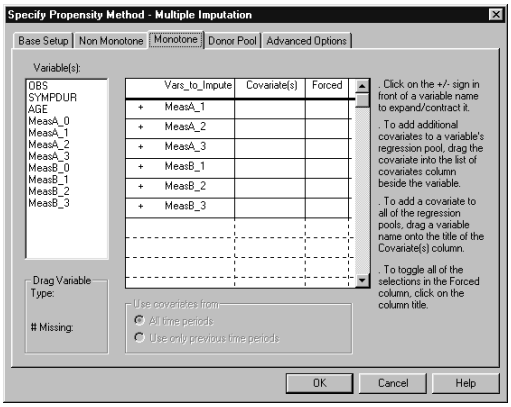
The list of covariates for each imputation variable will be made up of the variables specified as Fixed Covariates in the **Base Setup** tab, and all of the other imputation variables. Variables can be added and removed from this list of covariates by simply dragging and dropping the variable from the covariate list to the variables field, or vice versa. Even though a variable appears in the list of covariates for a particular imputation variable, it might not be used in the final model.

The program first sorts the variables so that the missing data pattern is as close as possible to monotone, and then, for each missing value in the imputation variable, the program works out which variables, from the total list of covariates, can be used for prediction.

By default, all of the covariates are forced into the model. If you uncheck a covariate, it will not be forced into the model, but will be retained as a possible covariate in the stepwise selection. Details of the models that were actually used to impute the missing values are included in the **Output log** that can be selected from the **View** menu of the Multiply-Imputed Data Pages. These data pages will be displayed after you have specified the imputation and pressed the **OK** button in the Specify Predictive Model window.

Monotone

Selecting the **Monotone** tab allows you to add or remove covariates from the logistic model used for imputing the monotone missing values in the data set. (These can be identified in the **Missing Data Pattern** mentioned earlier.)



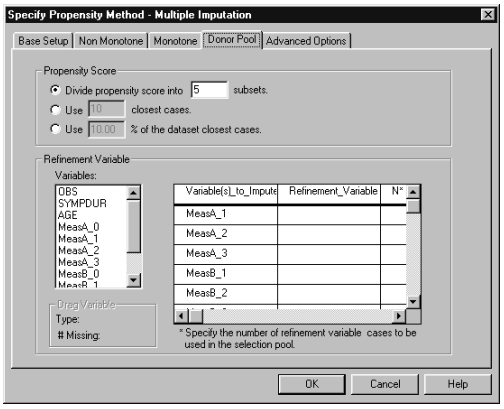
Again, you select the **+** or **-** signs to expand or contract the list of covariates for each imputation variable.

The list of covariates for each imputation variable will be made up of the variables specified as Fixed Covariates in the **Base Setup** tab, and all of the other imputation variables. Variables can be added and removed from this list by simply dragging and dropping the variable from the list of covariates, to the variables field, or vice versa. Even though a variable appears in the list of covariates for a particular imputation variable, it might not be used in the final model.

The program first sorts the variables so that the missing data pattern is as close as possible to monotone, and then uses only the variables that are to the left of the imputation variable as covariates. Details of the models that were actually used to impute the missing values are included in the **Output Log**.

Donor Pool

Selecting the **Donor pool** tab displays the Donor Pool page that allows more control over the random draw step in the analysis by allowing the user to define Propensity Score sub-classes.



The following options for defining the Propensity Score sub-classes are provided:

- **Divide propensity score into c subsets**. The default is 5.
- **Use c closest cases**. This option allows you to specify the number of closest cases that are to be included in the subset.
- **Use d% of the data set closest cases**. This option allows you to specify the number of cases as a

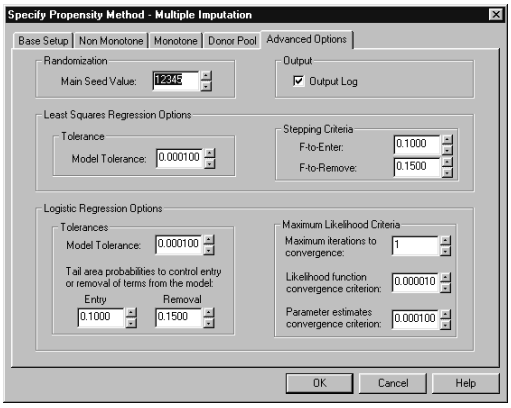
percentage.

NOTE: See *Defining Donor Pools Based on Propensity Scores* earlier in this manual.
You can use one Refinement Variable for each of the variables being imputed. Variables can be dragged from the **Variables** listbox to the **Refinement Variable** column. When you use a refinement variable, the program reduces the subset of cases included in the donor pool to include only cases that are close with respect to their values of the refinement variable.

You can also specify the number of refinement variable cases to be used in the donor pool. For this example, we will use all of the default settings in this tab.

Advanced Options

Selecting the **Advanced Options** tab displays the Advanced Options window that allows the user to control the settings for the imputation and the logistic regression.



Randomization

Main Seed Value

The **Main Seed Value** is used to perform the random selection within the propensity subsets. The default seed is 12345. If you set this field to blank, or set it to zero, then the clock time is used.

Output Log

The Output Log is a comprehensive list of regression equations etc. that have been calculated for the imputed variable(s).

Least Squares Regression

Tolerance

The value set in the **Tolerance** datafield controls numerical accuracy. The tolerance limit is used for matrix inversion to guard against singularity. No independent variable is used whose R^2 with other independent variables exceeds (1-Tolerance). You can adjust the tolerance using the scrolled datafield.

Stepping Criteria

Here you can select **F-to-Enter** and **F-to-Remove** values from the scrolled datafields, or enter your chosen value. If you wish to see more variables entered in the model, set the **F-to-Enter** value to a smaller value. The numerical value of **F-to-remove** should be chosen to be less than the **F-to-Enter** value.

Logistic Regression Options

Model Tolerance

Controls the numerical accuracy. Computations are performed in double precision. Use a value that is greater than .000001 but less than 1.0. The default is .0001.

Tail area probabilities to control entry or removal of terms from the model

Specifies the limits for the tail area probabilities (p-values) for the appropriate χ^2 and *F* values used to control the entry and removal of terms.

Entry

During forward stepping, the term with the smallest p-value less than the entry value is entered first. If no term in the model has a p-value less than this limit, then the term with the largest p-value greater than the removal value is removed.

Removal
During backward stepping, the term with the largest p-value greater than the removal value is removed first. Then any terms with entry p-values less than the entry limit are entered. Again, for the purposes of this example, we will run the analysis with the default settings.
Maximum Likelihood Criteria

Maximum iterations to convergence
Specifies the maximum number of iterations to maximize the likelihood function. The default is 10.

Likelihood function convergence criterion
Specifies the convergence criterion for the likelihood function. A relative improvement less than this value is considered no improvement. The default is .00001.

Parameter estimates convergence criterion
Specifies the convergence criterion for the parameter estimates. The default is .0001. When you are satisfied that you have specified your analysis correctly, click the **OK** button. The multiply-imputed datapages will be displayed, with the imputed values appearing in Red or Blue. Refer to *Analyzing Multiply-Imputed Data Sets* for further details of analyzing these data sets and combining the results.

Output

When you are satisfied that you have specified your analysis correctly, click the **OK** button. The multiply-imputed datapages will be displayed, with the imputed values appearing in Red or Blue. Refer to "Analyzing Multiple Imputed Data sets" (p. 49) for further details of analyzing these data sets and combining the results.

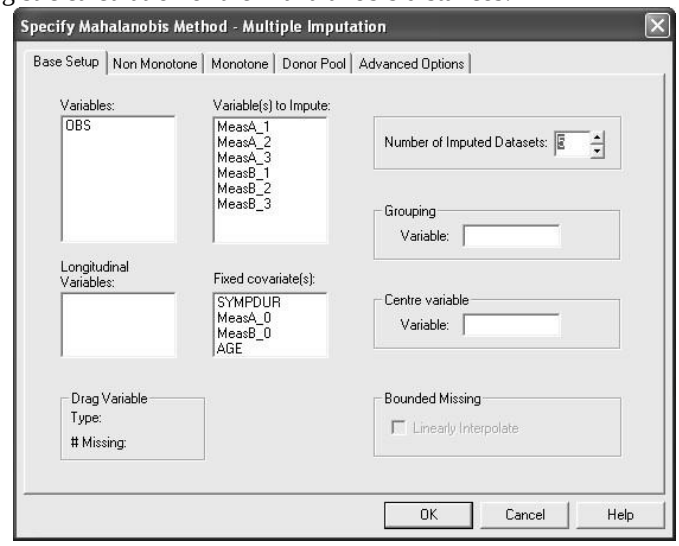
Mahalanobis Distance Matching Method — Example

We will now multiply impute all of the missing values in the data set using the Mahalanobis Distance Matching Method.

1. From the **Analyze** menu, select **Multiple Imputation** and **Mahalanobis Method**.
2. The **Specify Mahalanobis Method** window is displayed and is a tabbed (paged) window. The window opens with two pages or tabs: **Base Setup** and **Advanced Options**. As soon as you select a variable to be imputed, a **Non-Monotone** tab, a **Monotone** tab, and a **Donor Pool** tab are also displayed.

Base Setup

Selecting the **Base Setup** tab allows you specify which variables you want to impute, and which variables you want to use as covariates for the logistic calculation of the Mahalanobis distances.



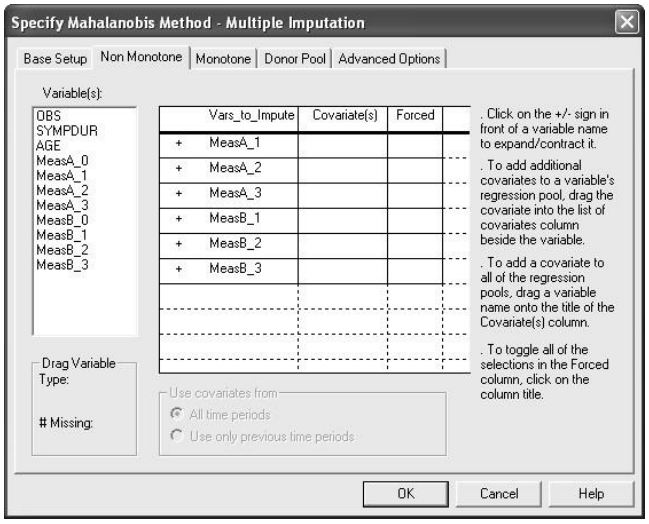
1. Drag-and-drop the variables **MeasA_1**, **MeasA_2**, **MeasA_3**, **MeasB_1**, **MeasB_2**, **MeasB_3** into the Variables to Impute field.
2. Drag and drop the variables **SYMPDUR**, **AGE**, **MeasA_0**, and **MeasB_0** into the Fixed Covariates field.
3. As there is no Grouping variable in this data set, we can leave this field blank.

If a centre variable is specified when calculating the Mahalanobis Distance, the covariance matrix used is a weighted average taken across the different levels of the Centre Variable, *(See Appendix F)*.

Non-Monotone

Selecting the **Non-monotone** tab allows you to add or remove covariates from the logistic model used for imputing the non-monotone missing values in the data set. (These can be identified in the **Missing Data Pattern** mentioned earlier in the Predictive Model example.)

You select the **+** or **-** signs to expand or contract the list of covariates for each imputation variable.



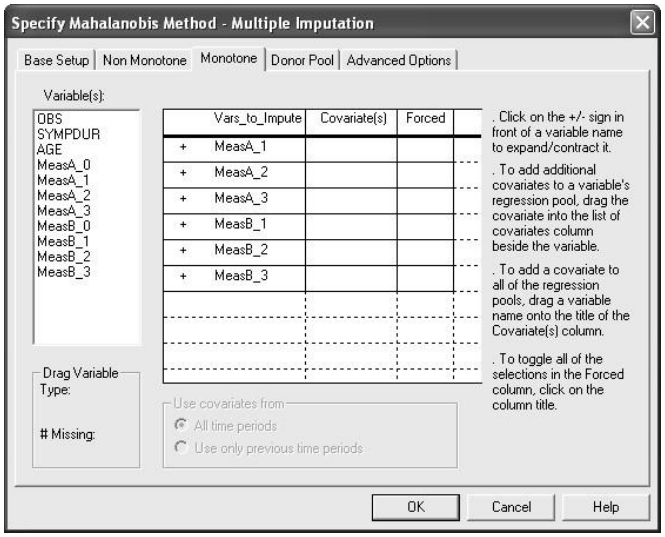
The list of covariates for each imputation variable will be made up of the variables specified as Fixed Covariates in the **Base Setup** tab, and all of the other imputation variables. Variables can be added and removed from this list of covariates by simply dragging and dropping the variable from the covariate list to the variables field, or vice versa. Even though a variable appears in the list of covariates for a particular imputation variable, it might not be used in the final model.

The program first sorts the variables so that the missing data pattern is as close as possible to monotone, and then, for each missing value in the imputation variable, the program works out which variables, from the total list of covariates, can be used for prediction.

By default, all of the covariates are forced into the model. If you uncheck a covariate, it will not be forced into the model, but will be retained as a possible covariate in the stepwise selection. Details of the models that were actually used to impute the missing values are included in the **Output log** that can be selected from the **View** menu of the Multiply-Imputed Data Pages. These data pages will be displayed after you have specified the imputation and pressed the **OK** button in the Specify Predictive Model window.

Monotone

Selecting the **Monotone** tab allows you to add or remove covariates used for calculating the Mahalanobis distances used for imputing the monotone missing values in the data set. (These can be identified in the **Missing Data Pattern** mentioned earlier.)



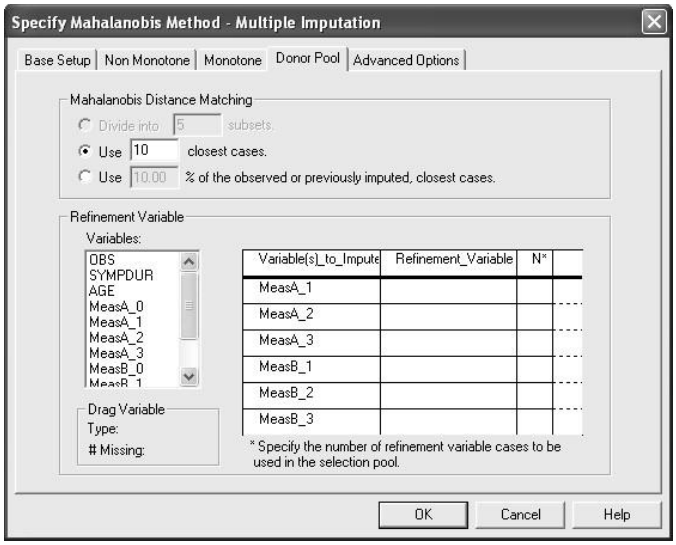
Again, you select the **+** or **-** signs to expand or contract the list of covariates for each imputation variable.

The list of covariates for each imputation variable will be made up of the variables specified as Fixed Covariates in the **Base Setup** tab, and all of the other imputation variables. Variables can be added and removed from this list by simply dragging and dropping the variable from the list of covariates, to the variables field, or vice versa.

The program first sorts the variables so that the missing data pattern is as close as possible to monotone, and then uses only the variables that are to the left of the imputation variable as covariates. Details of the models that were actually used to impute the missing values are included in the **Output Log**.

Donor Pool

Selecting the **Donor pool** tab displays the Donor Pool page that allows more control over the random draw step in the analysis by allowing the user to define Mahalanobis Distance sub-classes.



The following options for defining the Mahalanobis distance sub-classes are provided:

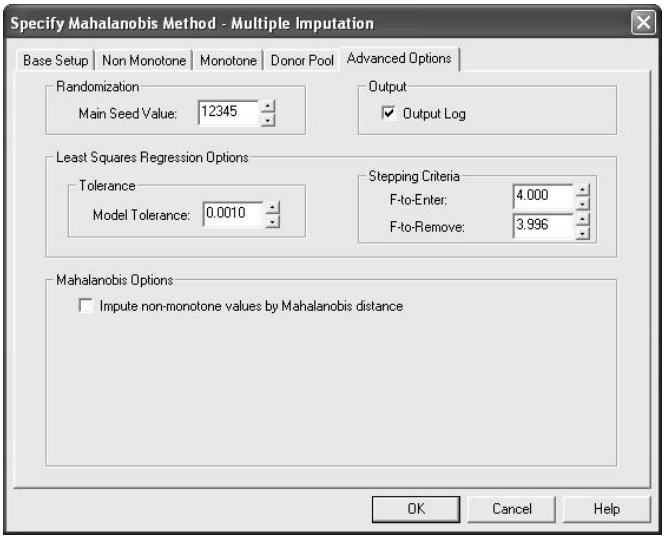
- **Use c closest cases**. This option allows you to specify the number of closest cases that are to be included in the subset.
- **Use d% of the data set closest cases**. This option allows you to specify the number of cases as a percentage.

NOTE: See *Defining Donor Pools Based on Mahalanobis distances* earlier in this manual.
You can use one Refinement Variable for each of the variables being imputed. Variables can be dragged from the **Variables** listbox to the **Refinement Variable** column. When you use a refinement variable, the program reduces the subset of cases included in the donor pool to include only cases that are close with respect to their values of the refinement variable.

You can also specify the number of refinement variable cases to be used in the donor pool. For this example, we will use all of the default settings in this tab.

Advanced Options

Selecting the **Advanced Options** tab displays the Advanced Options window that allows the user to control the settings for the imputation.



Randomization

Main Seed Value

The **Main Seed Value** is used to perform the random selection within the Mahalanobis distance subsets. The default seed is 12345. If you set this field to blank, or set it to zero, then the clock time is used.

Output Log

The Output Log is a comprehensive list of regression equations etc. that have been calculated for the imputed variable(s).

Least Squares Regression

Tolerance

The value set in the **Tolerance** datafield controls numerical accuracy. The tolerance limit is used for matrix inversion to guard against singularity. No independent variable is used whose R^2 with other independent variables exceeds (1-Tolerance). You can adjust the tolerance using the scrolled datafield.

Stepping Criteria

Here you can select **F-to-Enter** and **F-to-Remove** values from the scrolled datafields, or enter your chosen value. If you wish to see more variables entered in the model, set the **F-to-Enter** value to a smaller value. The numerical value of **F-to-remove** should be chosen to be less than the **F-to-Enter** value.

Mahalanobis Options

By selecting **Impute non-monotone values by Mahalanobis distance** this will force the system to use the Mahalanobis distance method to impute all values, whether they are in a Monotone or Non-Monotone pattern.

Output

When you are satisfied that you have specified your analysis correctly, click the **OK** button. The multiply-imputed datapages will be displayed, with the imputed values appearing in Red or Blue. Refer to "Analyzing Multiple Imputed Data sets" (p. 49) for further details of analyzing these data sets and combining the results.

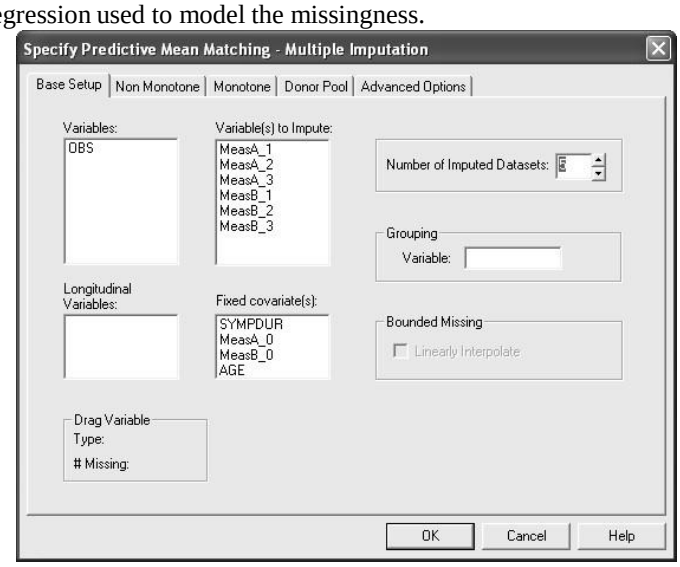
Predictive Mean Matching Method — Example

We will now multiply impute all of the missing values in the data set using the Predictive Mean Matching Method.

1. From the **Analyze** menu, select **Multiple Imputation** and **Predictive Mean Matching Method**.
2. The **Specify Predictive Mean Matching Method** window is displayed and is a tabbed (paged) window. The window opens with two pages or tabs: **Base Setup** and **Advanced Options**. As soon as you select a variable to be imputed, a **Non-Monotone** tab, a **Monotone** tab, and a **Donor Pool** tab are also displayed.

Base Setup

Selecting the **Base Setup** tab allows you specify which variables you want to impute, and which variables you want to use as covariates for the regression used to model the missingness.

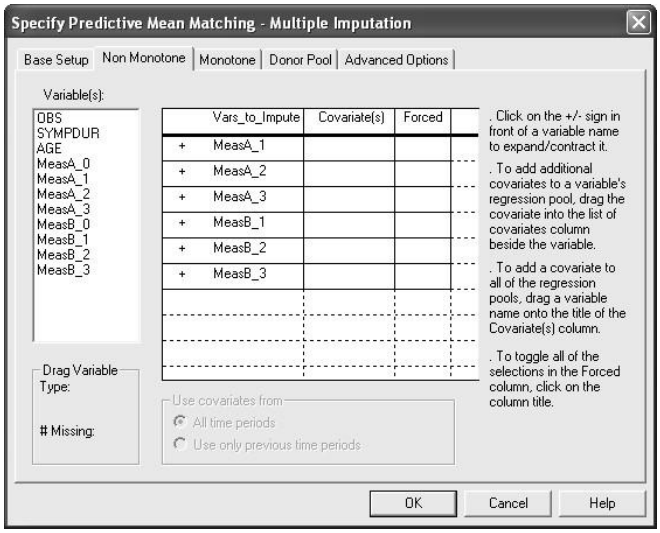


1. Drag-and-drop the variables **MeasA_1**, **MeasA_2**, **MeasA_3**, **MeasB_1**, **MeasB_2**, **MeasB_3** into the Variables to Impute field.
2. Drag and drop the variables **SYMPDUR**, **AGE**, **MeasA_0**, and **MeasB_0** into the Fixed Covariates field.
3. As there is no Grouping variable in this data set, we can leave this field blank.

Non-Monotone

Selecting the **Non-monotone** tab allows you to add or remove covariates from the regression model used for imputing the non-monotone missing values in the data set. (These can be identified in the **Missing Data Pattern** mentioned earlier in the Predictive Model example.)

You select the **+** or **-** signs to expand or contract the list of covariates for each imputation variable.



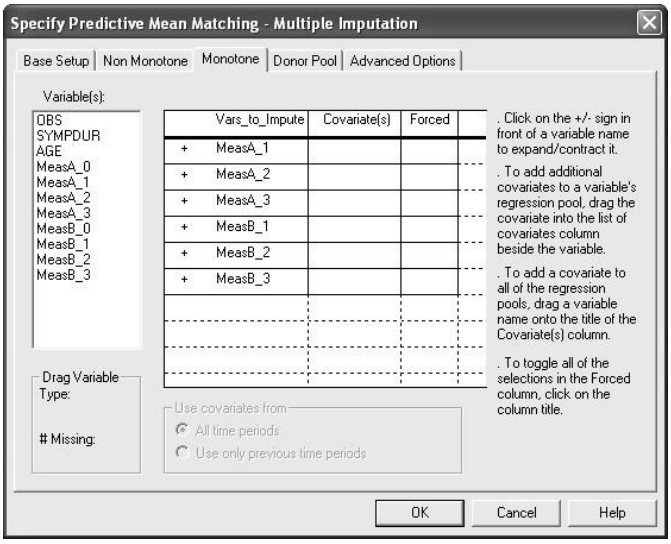
The list of covariates for each imputation variable will be made up of the variables specified as Fixed Covariates in the **Base Setup** tab, and all of the other imputation variables. Variables can be added and removed from this list of covariates by simply dragging and dropping the variable from the covariate list to the variables field, or vice versa. Even though a variable appears in the list of covariates for a particular imputation variable, it might not be used in the final model.

The program first sorts the variables so that the missing data pattern is as close as possible to monotone, and then, for each missing value in the imputation variable, the program works out which variables, from the total list of covariates, can be used for prediction.

By default, all of the covariates are forced into the model. If you uncheck a covariate, it will not be forced into the model, but will be retained as a possible covariate in the stepwise selection. Details of the models that were actually used to impute the missing values are included in the **Output log** that can be selected from the **View** menu of the Multiply-Imputed Data Pages. These data pages will be displayed after you have specified the imputation and pressed the **OK** button in the Specify Predictive Model window.

Monotone

Selecting the **Monotone** tab allows you to add or remove covariates from the model used for imputing the monotone missing values in the data set. (These can be identified in the **Missing Data Pattern** mentioned earlier.)



Again, you select the **+** or **-** signs to expand or contract the list of covariates for each imputation variable.

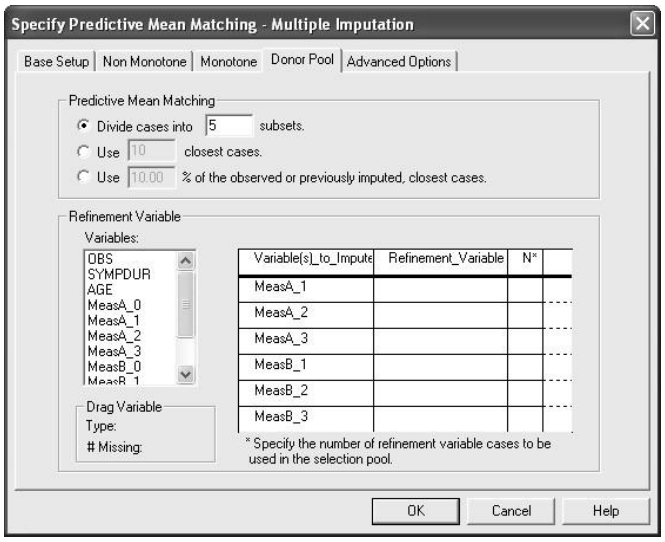
The list of covariates for each imputation variable will be made up of the variables specified as Fixed Covariates in the **Base Setup** tab, and all of the other imputation variables. Variables can be added and removed from this list by

simply dragging and dropping the variable from the list of covariates, to the variables field, or vice versa. Even though a variable appears in the list of covariates for a particular imputation variable, it might not be used in the final model.

The program first sorts the variables so that the missing data pattern is as close as possible to monotone, and then uses only the variables that are to the left of the imputation variable as covariates. Details of the models that were actually used to impute the missing values are included in the **Output Log**.

Donor Pool

Selecting the **Donor pool** tab displays the Donor Pool page that allows more control over the random draw step in the analysis by allowing the user to define Propensity Score sub-classes.



The following options for defining the Propensity Score sub-classes are provided:

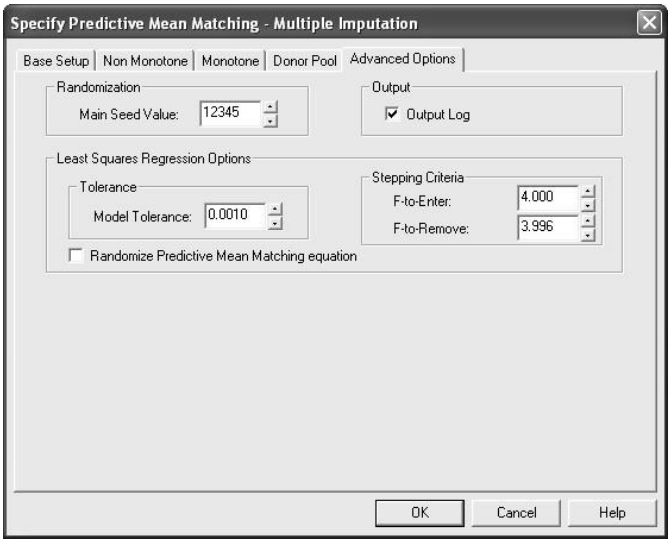
- **Divide predicted values into c subsets**. The default is 5.
- **Use c closest cases**. This option allows you to specify the number of closest cases that are to be included in the subset.
- **Use d% of the data set closest cases**. This option allows you to specify the number of cases as a percentage.

NOTE: See *Defining Donor Pools Based on Predicted Values* earlier in this manual.
You can use one Refinement Variable for each of the variables being imputed. Variables can be dragged from the **Variables** listbox to the **Refinement Variable** column. When you use a refinement variable, the program reduces the subset of cases included in the donor pool to include only cases that are close with respect to their values of the refinement variable.

You can also specify the number of refinement variable cases to be used in the donor pool. For this example, we will use all of the default settings in this tab.

Advanced Options

Selecting the **Advanced Options** tab displays the Advanced Options window that allows the user to control the settings for the imputation and the logistic regression.



Randomization

Main Seed Value

The **Main Seed Value** is used to perform the random selection within the predicted value subsets. The default seed is 12345. If you set this field to blank, or set it to zero, then the clock time is used.

Output Log

The Output Log is a comprehensive list of regression equations etc. that have been calculated for the imputed variable(s).

Least Squares Regression

Tolerance

The value set in the **Tolerance** datafield controls numerical accuracy. The tolerance limit is used for matrix inversion to guard against singularity. No independent variable is used whose R^2 with other independent variables exceeds (1-Tolerance). You can adjust the tolerance using the scrolled datafield.

Stepping Criteria

Here you can select **F-to-Enter** and **F-to-Remove** values from the scrolled datafields, or enter your chosen value. If you wish to see more variables entered in the model, set the **F-to-Enter** value to a smaller value. The numerical value of **F-to-remove** should be chosen to be less than the **F-to-Enter** value.

Randomize Predictive Mean Matching equation

If this option is selected then the same approach as in the Predictive Model based method is used to randomly draw the coefficients for the prediction equation from the posterior distribution of the estimated coefficients.

Output

When you are satisfied that you have specified your analysis correctly, click the **OK** button. The multiply-imputed datapages will be displayed, with the imputed values appearing in Red or Blue. Refer to "Analyzing Multiple Imputed Data sets" (p. 49) for further details of analyzing these data sets and combining the results.

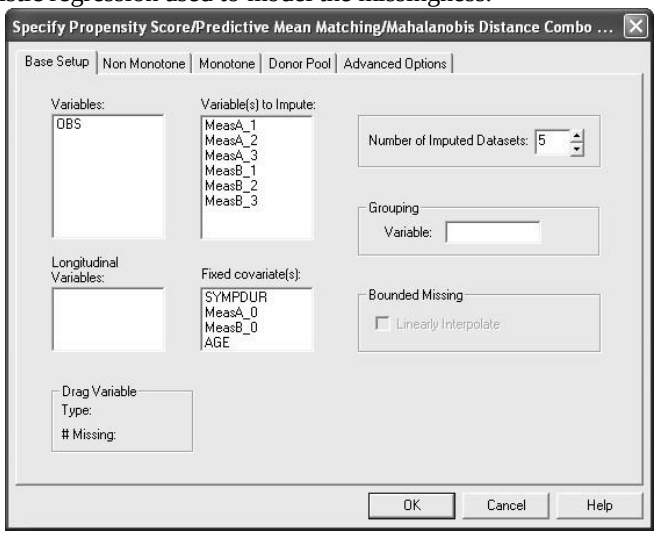
Propensity Score/Predictive Mean/Mahalanobis Distance Combination Method — Example

We will now multiply impute all of the missing values in the data set using the Combination Method.

1. From the **Analyze** menu, select **Multiple Imputation** and **Propensity/Predictive/Mahalanobis Combo Method**.
2. The **Specify Propensity Score/Predictive Mean Matching/Mahalanobis Distance Combo Method** window is displayed and is a tabbed (paged) window. The window opens with two pages or tabs: **Base Setup** and **Advanced Options**. As soon as you select a variable to be imputed, a **Non-Monotone** tab, a **Monotone** tab, and a **Donor Pool** tab are also displayed.

Base Setup

Selecting the **Base Setup** tab allows you specify which variables you want to impute, and which variables you want to use as covariates for the logistic regression used to model the missingness.

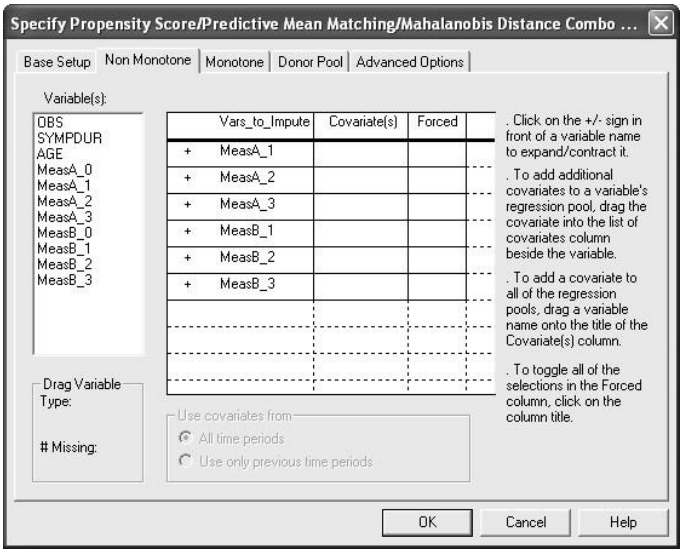


1. Drag-and-drop the variables **MeasA_1**, **MeasA_2**, **MeasA_3**, **MeasB_1**, **MeasB_2**, **MeasB_3** into the Variables to Impute field.
2. Drag and drop the variables **SYMPDUR**, **AGE**, **MeasA_0**, and **MeasB_0** into the Fixed Covariates field.
3. As there is no Grouping variable in this data set, we can leave this field blank.

Non-Monotone

Selecting the **Non-monotone** tab allows you to add or remove covariates from the model used for imputing the non-monotone missing values in the data set. (These can be identified in the **Missing Data Pattern** mentioned earlier in the Predictive Model example.)

You select the **+** or **-** signs to expand or contract the list of covariates for each imputation variable.



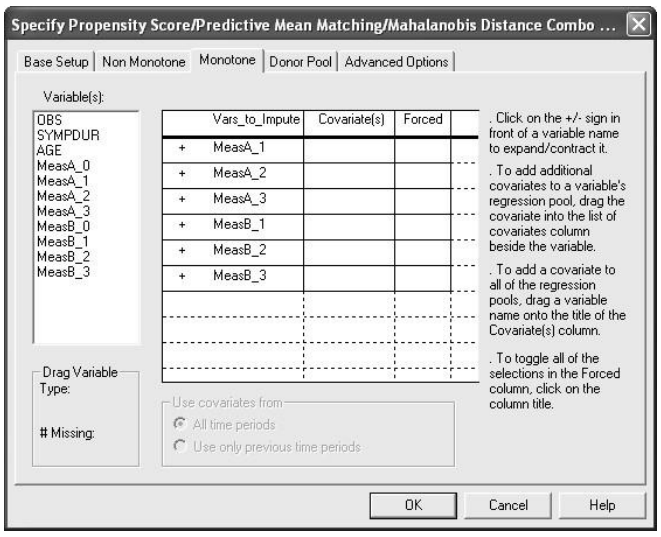
The list of covariates for each imputation variable will be made up of the variables specified as Fixed Covariates in the **Base Setup** tab, and all of the other imputation variables. Variables can be added and removed from this list of covariates by simply dragging and dropping the variable from the covariate list to the variables field, or vice versa. Even though a variable appears in the list of covariates for a particular imputation variable, it might not be used in the final model.

The program first sorts the variables so that the missing data pattern is as close as possible to monotone, and then, for each missing value in the imputation variable, the program works out which variables, from the total list of covariates, can be used for prediction.

By default, all of the covariates are forced into the model. If you uncheck a covariate, it will not be forced into the model, but will be retained as a possible covariate in the stepwise selection. Details of the models that were actually used to impute the missing values are included in the **Output log** that can be selected from the **View** menu of the Multiply-Imputed Data Pages. These data pages will be displayed after you have specified the imputation and pressed the **OK** button in the Specify Predictive Model window.

Monotone

Selecting the **Monotone** tab allows you to add or remove covariates from the logistic model used for imputing the monotone missing values in the data set. (These can be identified in the **Missing Data Pattern** mentioned earlier.)



Again, you select the **+** or **-** signs to expand or contract the list of covariates for each imputation variable.

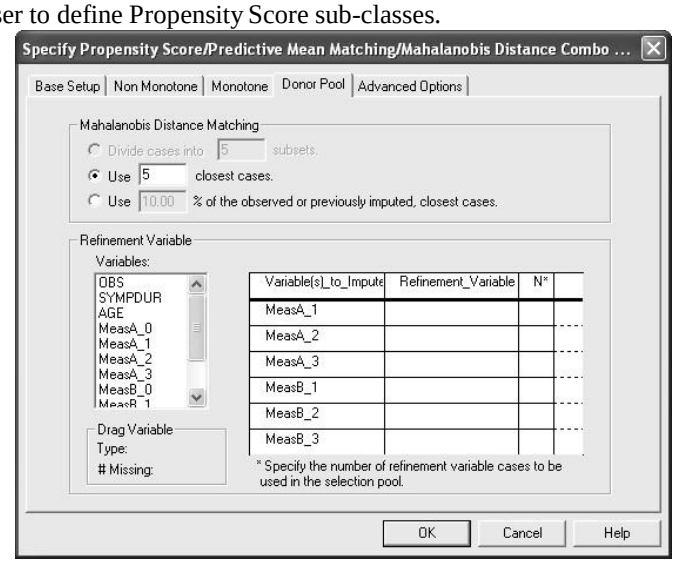
The list of covariates for each imputation variable will be made up of the variables specified as Fixed Covariates in the **Base Setup** tab, and all of the other imputation variables. Variables can be added and removed from this list by simply dragging and dropping the variable from the list of covariates, to the variables field, or vice versa. Even

though a variable appears in the list of covariates for a particular imputation variable, it might not be used in the final model.

The program first sorts the variables so that the missing data pattern is as close as possible to monotone, and then uses only the variables that are to the left of the imputation variable as covariates. Details of the models that were actually used to impute the missing values are included in the **Output Log**.

Donor Pool

Selecting the **Donor pool** tab displays the Donor Pool page that allows more control over the random draw step in the analysis by allowing the user to define Propensity Score sub-classes.



The following options for defining the Propensity Score sub-classes are provided:

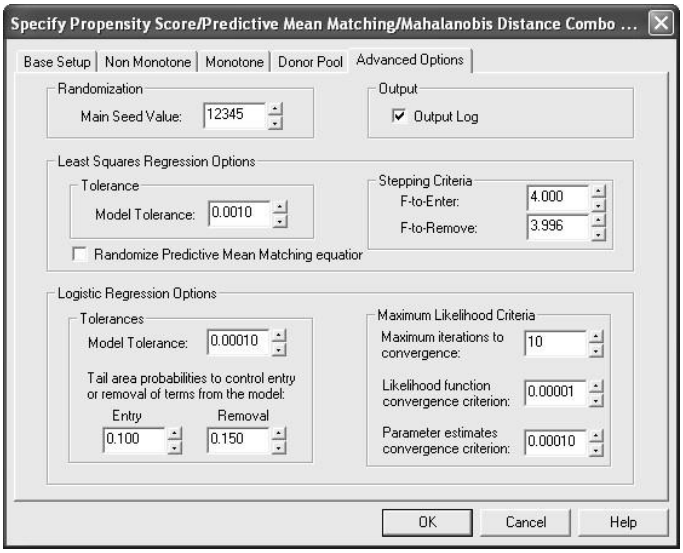
- **Use c closest cases**. This option allows you to specify the number of closest cases that are to be included in the subset.
- **Use d% of the data set closest cases**. This option allows you to specify the number of cases as a percentage.

NOTE: See *Defining Donor Pools Based on Mahalanobis Distances* earlier in this manual.
You can use one Refinement Variable for each of the variables being imputed. Variables can be dragged from the **Variables** listbox to the **Refinement Variable** column. When you use a refinement variable, the program reduces the subset of cases included in the donor pool to include only cases that are close with respect to their values of the refinement variable.

You can also specify the number of refinement variable cases to be used in the donor pool. For this example, we will use all of the default settings in this tab.

Advanced Options

Selecting the **Advanced Options** tab displays the Advanced Options window that allows the user to control the settings for the imputation.



Randomization

Main Seed Value

The **Main Seed Value** is used to perform the random selection within the propensity subsets. The default seed is 12345. If you set this field to blank, or set it to zero, then the clock time is used.

Output Log

The Output Log is a comprehensive list of regression equations etc. that have been calculated for the imputed variable(s).

Least Squares Regression

Tolerance

The value set in the **Tolerance** datafield controls numerical accuracy. The tolerance limit is used for matrix inversion to guard against singularity. No independent variable is used whose R^2 with other independent variables exceeds (1-Tolerance). You can adjust the tolerance using the scrolled datafield.

Stepping Criteria

Here you can select **F-to-Enter** and **F-to-Remove** values from the scrolled datafields, or enter your chosen value. If you wish to see more variables entered in the model, set the **F-to-Enter** value to a smaller value. The numerical value of **F-to-remove** should be chosen to be less than the **F-to-Enter** value.

Randomize Predictive Mean Matching equation

If this option is selected then the same approach as in the Predictive Model based method is used to randomly draw the coefficients for the prediction equation from the posterior distribution of the estimated coefficients.

Logistic Regression Options

Model Tolerance

Controls the numerical accuracy. Computations are performed in double precision. Use a value that is greater than .000001 but less than 1.0. The default is .0001.

Tail area probabilities to control entry or removal of terms from the model

Specifies the limits for the tail area probabilities (p-values) for the appropriate χ^2 and F values used to control the entry and removal of terms.

Entry

During forward stepping, the term with the smallest p-value less than the entry value is entered first. If no term in the model has a p-value less than this limit, then the term with the largest p-value greater than the removal value is removed.

Removal

During backward stepping, the term with the largest p-value greater than the removal value is removed first. Then any terms with entry p-values less than the entry limit are entered. Again, for the purposes of this example, we will run

the analysis with the default settings.

Maximum Likelihood Criteria

Maximum iterations to convergence
Specifies the maximum number of iterations to maximize the likelihood function. The default is 10.

Likelihood function convergence criterion
Specifies the convergence criterion for the likelihood function. A relative improvement less than this value is considered no improvement. The default is .00001.

Parameter estimates convergence criterion
Specifies the convergence criterion for the parameter estimates. The default is .0001. When you are satisfied that you have specified your analysis correctly, click the **OK** button. The multiply-imputed datapages will be displayed, with the imputed values appearing in Red or Blue. Refer to *Analyzing Multiply-Imputed Data Sets* for further details of analyzing these data sets and combining the results.

Output

When you are satisfied that you have specified your analysis correctly, click the **OK** button. The multiply-imputed datapages will be displayed, with the imputed values appearing in Red or Blue. Refer to "Analyzing Multiple Imputed Data sets" (p. 49) for further details of analyzing these data sets and combining the results.

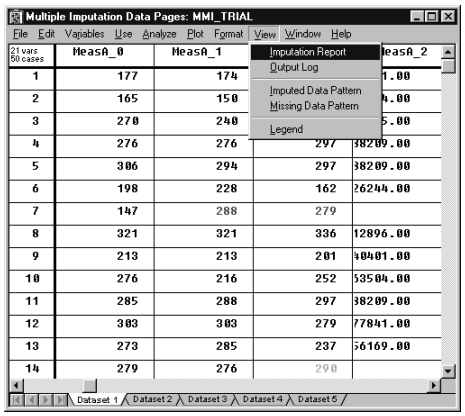
Multiple Imputation Output

The Multiple Imputation output, either Propensity Score or The Predictive Model Based Method comprises five (default value) Multiple Imputation Data Pages. From the **View** menu of the Data Pages you can select either: **Imputation Report**, **Output Log**, **Imputed Data Pattern**, or **Missing Data Pattern**.

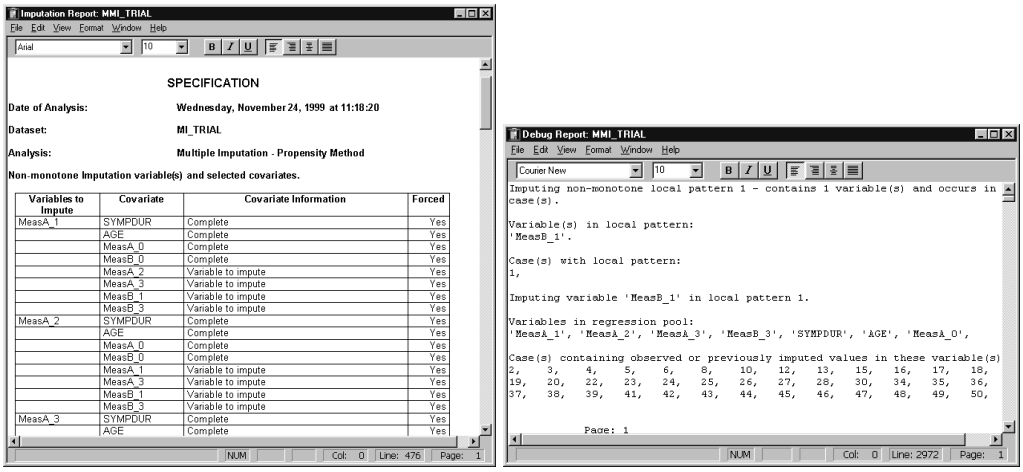
When other analyses are performed from the **Analyze** menu of a data page (see the example *Analyzing Multiply-Imputed Data Sets* later in this manual), a **Combined** tab is added to the data page tabs. Selecting this tab displays the combined statistics for these data pages. The combined statistics that are displayed are given *in Appendix B – Combined Statistics.*

Data Pages

The Multiple Imputation output displays five data pages with the imputed values shown in a color that contrasts with the observed values. These five pages of completed data results are displayed and allow the user to examine how the combined results are calculated. The first data page (Page 1) for the above example is shown below:



From the **View** menu, you can select **Imputation Report** and **Output Log** (examples of both are shown below) or **Imputed Data Pattern** and **Missing Data Pattern**.



Imputation Report: MMI_TRIAL

File Edit View Format Window Help

SPECIFICATION

Date of Analysis: Wednesday, November 24, 1999 at 11:18:20

Dataset: MI_TRIAL

Analysis: Multiple Imputation - Propensity Method

Non-monotone Imputation variable(s) and selected covariates.

Variables to Impute	Covariate	Covariate Information	Forced
MeasA_1	SYMPDUR	Complete	Yes
	AGE	Complete	Yes
	MeasA_0	Complete	Yes
	MeasB_0	Complete	Yes
	MeasA_2	Variable to impute	Yes
	MeasA_3	Variable to impute	Yes
	MeasB_1	Variable to impute	Yes
	MeasB_3	Variable to impute	Yes
MeasA_2	SYMPDUR	Complete	Yes
	AGE	Complete	Yes
	MeasA_0	Complete	Yes
	MeasB_0	Complete	Yes
	MeasA_1	Variable to impute	Yes
	MeasA_3	Variable to impute	Yes
	MeasB_1	Variable to impute	Yes
	MeasB_3	Variable to impute	Yes
MeasA_3	SYMPDUR	Complete	Yes
	AGE	Complete	Yes

NUM Col: 0 Line: 476 Page: 1

Debug Report: MMI_TRIAL

File Edit View Format Window Help

```
Imputing non-monotone local pattern 1 - contains 1 variable(s) and occurs in
case(s).

Variable(s) in local pattern:
'MeasB_1'.

Case(s) with local pattern:
1,

Imputing variable 'MeasB_1' in local pattern 1.

Variables in regression pool:
'MeasA_1', 'MeasA_2', 'MeasA_3', 'MeasB_3', 'SYMPDUR', 'AGE', 'MeasA_0',

Case(s) containing observed or previously imputed values in these variable(s)
2,    3,    4,    5,    6,    8,    10,   12,   13,   15,   16,   17,   18,
19,   20,   22,   23,   24,   25,   26,   27,   28,   30,   34,   35,   36,
37,   38,   39,   41,   42,   43,   44,   45,   46,   47,   48,   49,   50,

                    Page: 1
```

NUM Col: 0 Line: 2972 Page: 1

The Imputation Report and the Output Log (shown in part above), summarize the results of the logistic regression, the ordinary regression, and the settings used for the multiple imputation.

Imputation Report

The imputation report contains a summary of the parameters that were chosen for the Multiple Imputation. For example, the seed value that was used for the random selection, the number of imputations that were performed etc, are all reported. The report shows:

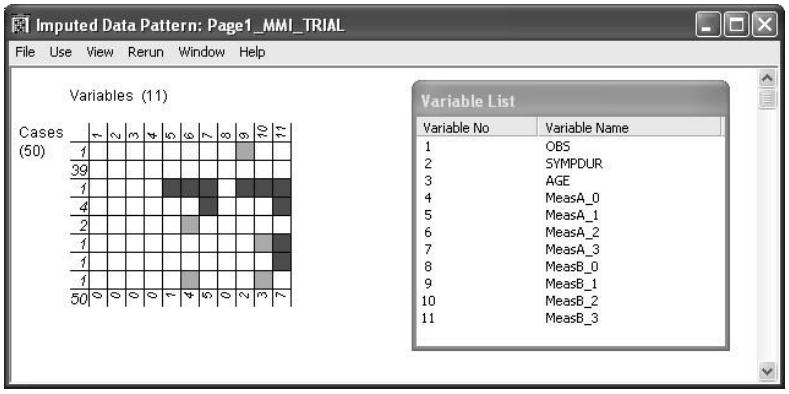
- An overview of the Multiple Imputation parameters.
- In the specification section there are tables of the variables and selected covariates for non- monotone and monotone patterns, number of imputed pages, random seed etc.
- Diagnostic information that can be used to judge the quality and validity of the generated imputations. The options chosen for the least squares and logistic regression options as well as sub-classing of propensity scores.
- The diagnostic section also gives a detailed breakdown of the number of cases available initially and numbers excluded for various reasons. Further conclusions about the statistical analysis can be drawn from the combined results (see *Analyzing Multiply- Imputed Data Sets* later in this manual).

Output Log

The Output Log provides details of the regressions carried out for all the imputed values on the imputed data sets. Information is given for the variables and cases involved in local patterns and the variables and cases involved in the regressions used. For propensity method propensity scores are given. For predictive model the equations used to estimate and generate the imputed values along with their error terms are given. For the Mahalanobis distance method the Mahalanobis distances are listed. For the predictive mean matching method the predicted means are given. For the combination method the propensity scores and predicted values are given followed by the Mahalanobis distances calculated using the propensity scores and predicted values as covariates.

Imputed Data Pattern and Missing Data Pattern windows

The **Missing Data Pattern** window can be selected from the **View** menu of the your datasheet before the imputation is performed. You can also display a colored legend from the **View** menu that identifies missing data and data that is present in the data set. The **Imputed Data Pattern** window can be selected from the **View** menu of the datasheet after the imputation is performed. You can also display a colored legend from the **View** menu that identifies Monotone and Non-monotone patterns. Example of the collapsed imputed data pattern is given below:



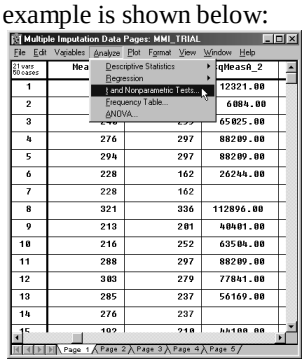
Analyzing Multiply-Imputed Data Sets — Example

This section presents a simple example of analyzing multiply-imputed data sets. It will show how the results of the repeated imputations can be combined to create one repeated imputation inference. For reference see *Appendix A – Analyzing Multiply-Imputed Data Sets*, and *Appendix F*, [1] and [2].

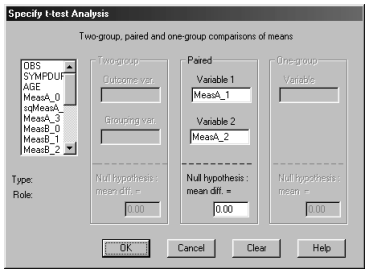
After you have performed a Multiple Imputation on your data set, you will have M complete data sets, each of which can be analyzed using standard complete-data statistical methods.

If you select **Descriptive Statistics**, **Regression**, **t-test**, **Frequency Table** from the **Analyze** menu of any data page, the analysis will be performed on all 5 data sets. The analysis generates 5 pages of output, one corresponding to each of the imputed data sets, and a Combined page that gives the overall set of results. The tabs at the bottom of the page allow you to display each data set.

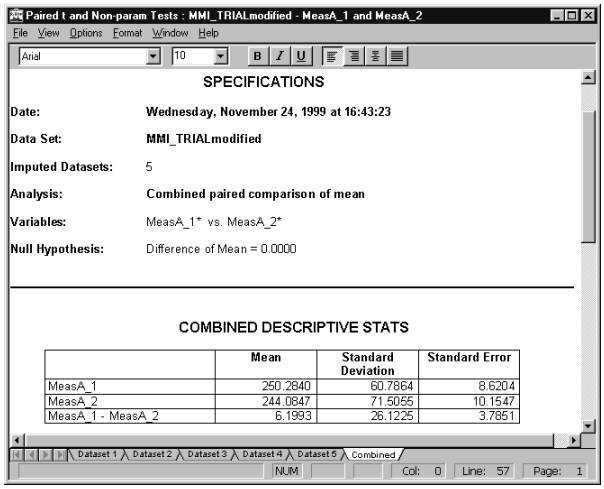
This example uses the imputation results from the data set **MI_TRIAL.MDD** that was used in the Propensity Score example earlier. Part of data page 1 for that example is shown below:



1. From the data page **Analyze** menu select **t- and Nonparametric Tests** to display the Specify t-test Analysis window.



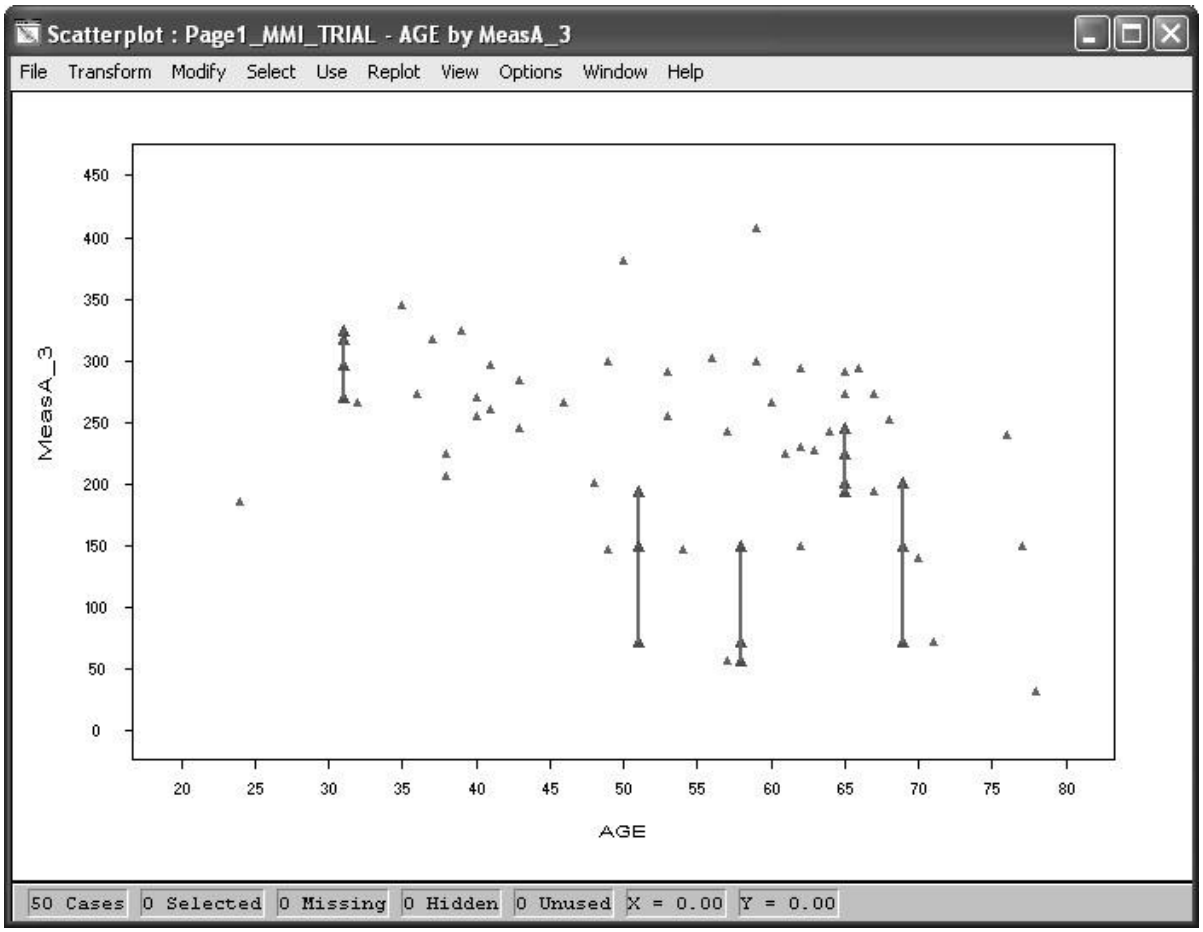
2. Drag and drop the variables **MeasA_1** and **MeasA_2** to the **Variable 1** and **Variable 2** datafields respectively.
3. Press the **OK** button to display the data pages, then press the **Combine** tab to display the combined statistics from the five imputed data pages as shown below.



The statistics that are calculated for each analysis selected from the **Analyze** menu and displayed in the Combined page are given in *Appendix B – Combined Statistics*.

Scatter Plots of Imputed Data

After imputation it is desirable to see the spread of the values that were imputed. This can be seen by using the **Scatterplot** option in the **Plot** menu. This can be done from any of the imputed datapages. Then, from the **View** menu, select **Multiple Imputation > Show all points** and **Draw lines**. This will then plot all imputed values and also include a line to indicate the range from the lowest to the highest impute value.



Glossary

DEFINITIONS

Bounded missing	A missing value (in a longitudinal variable) which has at least one observed value before and at least one observed value after the period for which it is missing.
Covariate	A variable which is selected as covariate for all selected variables to be imputed. Except for discriminant imputation, this variable is an independent variable in the corresponding regression model.
Fixed Covariate	A variable which is selected as covariate for all selected variables to be imputed.
Forced Covariate	A covariate that has been forced into a regression model, i.e. will not be removed from the model during stepping.
Hot-deck imputation	A method of imputation in which missing values are replaced with values taken from matching respondents (i.e. respondents that are similar with respect to variables observed for both).
Imputation	A procedure whereby missing values in a data set are filled-in with plausible estimates, to produce a complete data set which can then be analyzed using complete-data inferential methods.
Intent-to-treat	Intent-to-treat (IT) analysis dictates that all cases, both complete and incomplete, be included in any analyses, and treatment effects should be measured with subjects assigned to the treatment to which they were randomized, rather than to the treatment actually received.
Last value carried forward	A method of imputation for replacing missing values in longitudinal studies using the last observed value.
Longitudinal variable	A variable that is made up of a set of repeated measurements over time.
Mean imputation	The sample mean of a variable is used to replace any missing data for that variable. This mean can be an overall mean of all the cases, or a within group or class mean.
Multiple imputation	Each missing value is replaced by two or more (M) plausible estimates in order to create M complete data sets.
Possible Covariate	A covariate that has not been forced into a regression model, and so can be entered or removed during stepping.
Propensity score	Is the conditional probability of missingness computed from a vector of observed covariates.
Random imputation	A respondent is chosen at random from the total respondent sample for a variable, and the missing value for a non-respondent is replaced by the respondent's value.
Combine	The procedure for combining the set of M results into one overall set of results.
Imputation variable	A variable that has values that need to be imputed.

Appendix A: Analyzing Multiply-Imputed Data sets

ESTIMATED PARAMETERS

Definitions of Estimated Parameters

The following shows how M complete data analyses can be combined to create one repeated imputation inference. See Rubin and Schenker, 1991. Multiple Imputation in Health-Care Data Bases: An Overview and Some Applications, Statistics in Medicine, 10, 585-598, and Rubin D.B. (1987), Multiple Imputation for Non-response in Surveys, New York: John Wiley.

For each of the M complete data sets, let $\hat{\Theta}_m$, m = 1, ... , M, be M complete-data estimates for a parameter Θ, and U_m, m = 1, ... , M, be their associated variances.

Combined Estimate of Parameter

The combined estimate of any multi-dimensional parameter of interest Θ, for a particular variable is simply the mean of the estimates from each of the M imputed data sets. For example, the combined estimate of the mean for a specific group, or a particular regression coefficient in a model, is simply the mean of the estimates for that parameter across the M computed data sets:

$$\overline{\Theta} = \sum \hat{\Theta}_m / M$$

The general formula for combining point estimates:

$$\overline{Q}_m = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i$$

In some cases, point estimates are combined in a slightly different way:

- Standard deviation
- Serial correlation and Pearson *r*.

Where *m* = number of imputations and $\hat{Q}_i$ corresponds to the point estimate calculated from the i[th] datasheet.

Pooled Correlation = $\frac{\exp(2*\bar{z})-1}{\exp(2*\bar{z})+1}$, $\bar{z} = \frac{1}{m}\sum_{i=1}^{m} z_i$, and $z_i = 0.5*\ln\left(\frac{1+\hat{Q}_i}{1-\hat{Q}_i}\right)$

$\hat{Q}_i$ = Correlation for the i[th] imputed data set.

The pooled Standard Deviation = $\sqrt{\frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i}$ where *m* = number of imputations and $\hat{Q}_i$ corresponds to the variance calculated from the i[th] datasheet.

Standard Errors and Confidence Intervals

To estimate the variance of the combined parameter estimate, we combine the corresponding variance that is estimated from the combined parameter estimates from within each imputed data set, with the variability of the estimate across *m* imputed data sets. The standard error of a combined parameter estimate can be found by taking the square root of the variance of a combined parameter estimate.

The pooled standard error of a point-estimate = SE $\overline{Q}_m = \sqrt{T_m}$

$T_m = \overline{U}_m + (1 + m^{-1}) * B_m$ where $\overline{U}_m = \frac{1}{m}\sum_{i=1}^{m} U_i$ is the within imputation variance, where

U_i = [the standard error of the point-estimate from the i[th] data set][2] and:

$B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_i - \overline{Q}_m\right)^2$ is the between imputation variance, where $\hat{Q}_i$ = corresponding point-estimate calculated from the i[th] data set.

The pooled confidence interval for the point-estimate = $\overline{Q}_m \pm t\left(\tilde{v}_m, 1-\frac{\alpha}{2}\right) * SE\overline{Q}_m$

where α corresponds to a (1-α)100% C.I. and SE $\overline{Q}_m$ is the pooled standard error of the point-estimate as shown above.

$$\tilde{v}_m = v_{com}\left(\frac{1-\left[\hat{\gamma}_m/(m+1)\right]}{1-\hat{\gamma}_m} + \frac{v_{com}}{v_m}\right)^{-1}$$

See John Barnard and Donald B. Rubin, *Biometrika,* Small sample degrees of freedom with multiple imputation, December 1999, Volume 86, No. 4.

where v_{com}= degrees of freedom used in case of complete data and where:

$$\hat{\gamma}_m = \frac{\left(1+m^{-1}\right) * B_m}{T_m}$$

and:

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_i - \overline{Q}_m\right)^2 \text{ and } T_m = \overline{U}_m + \left(1+m^{-1}\right) * B_m$$

and:

$v_m = (m-1)*\left(1+r_m^{-1}\right)^2$ and $r_m = \frac{\left(1+m^{-1}\right) * B_m}{\overline{U}_m}$ and $\overline{U}_m = \frac{1}{m}\sum_{i=1}^{m} U_i$ where

U_i = [the standard error of the point-estimate from the i[th] data set]^2.

Appendix B: Combined Statistics

STATISTICS

Combined Statistics for Imputed Data sets

Pressing the **Combined** tab in a data page displays the statistics computed using the results of the **M** analyses. Each statistic is first combined across the M results. Each displayed statistic is then followed by a series of diagnostics useful in assessing the effect of the missing data on the statistical result.

For example, if the mean is computed in a Descriptive Statistics output, the associated combined statistics for the mean include:

- The average of the **M** computed means, its total variance T_m, and its total standard error $\sqrt{T_m}$.

The Diagnostics include:

- The between imputation variance (B_m), the between imputation standard error (sqrt(B_m)), the relative increase in variance due to missing data(r_m), sqrt(r_m), and the fraction of information missing due to missing data γ.

The statistics that are combined for each analysis are listed below.

Descriptive Statistics

- Mean, C.I. for mean, Standard deviation, Standard error of mean, Variance
- Coefficient of variation, Skewness, Kurtosis, Median, Quartiles
- Interquartile range, Proportion
- Serial Correlation

t and Non-parametric Tests

Descriptive Statistics

- Means
- Standard deviations
- Standard errors of the means
- Confidence intervals for the means

Two-group

- Pooled Variance t-test including t- value, df and p-values

Paired

- Matched t-test including t- value, df and p-value

One-group

- Pooled Variance t-test including t- value, df and p-value

Frequency Table

Tables

- Row percentages
- Column percentages
- Total percentages

Associated Measures

- Odds-ratio including ln Odds ratio
- Kappa statistic
- Cramer's V
- Phi

Test Statistic

- Likelihood ratio chi-square

Multiple Regression

Regression Statistics

- Square root of Residual Mean Square
- Multiple Correlation
- Multiple Correlation Squared

Analysis of Variance

- F-Value
- p-value

Regression coefficients

- Partial Correlation
- Estimate of coefficient
- Standard error of coefficient
- Standardized coefficient
- t-value of coefficient
- Confidence interval of coefficient
- Pooled Multiple Linear regression Equation

Appendix C: Multiple Imputation – Predictive Model Based Method

COMPLETELY OBSERVED COVARIATES
INCOMPLETELY OBSERVED COVARIATES

Definition of Methods

The following gives a detailed explanation of the methods used to analyze situations with completely and incompletely observed covariates for Linear Regression Based Multiple Imputation.

Completely observed covariates

Let *y* be one imputation variable and let $x_1,\ldots, x_p$ be the fully observed covariates for y. Let Y_{obs} and Y_{mis} be the observed and missing data for y, respectively. Let *X* be the data matrix for $x_1,\ldots,x_p$. The first column of *X* consists of 1's to adjust for the intercept term and the second until the last column contains the observations for $x_1,\ldots, x_p$. Let X_{obs} and X_{mis} be the rows of *X* corresponding to Y_{obs} and Y_{mis}, respectively. The underlying statistical model of linear regression imputation is given by:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \text{, where } \varepsilon \sim N\left(0, \sigma^2\right)$$

Let q be equal to p+1. The parameter q equals the number of regression coefficients including the intercept. Each imputation Y_{mis}* for Y_{mis} is independently generated in the following steps:

1. Let $\hat{\beta}$ and $\hat{\sigma}^2$ be the least squares estimators of $\beta = \left(\beta_0, \beta_1, \ldots, \beta_p\right)$ and of σ^2 from Y_{obs} and X_{obs}.

   Let *V* be the inverse of the matrix $[X_{obs}]^T[X_{obs}]$, and $V^{1/2}$ be a square root of *V* that can be obtained via the Choleski decomposition of *V*.

   Let *P* be the matrix of eigenvectors of *V* and Λ be the diagonal matrix with Λ_{ii} equal to the eigenvalue of *V* corresponding to the eigenvector of *V* given by the i^{th} column in *P*.

   The square root $V^{1/2}$ of *V* is then given by $V^{1/2} = P\Lambda^{1/2}$ with $\Lambda^{1/2}$ with the diagonal matrix containing the square roots of Λ_{ii} as its diagonal elements.
2. Draw a $\chi^2_{n_{obs}-q}$ random variable g.
3. Let $\sigma^{*2} = \hat{\sigma}^2\left(n_{obs} - q\right)/g$.
4. Draw *q* independent random variables $Z_1,\ldots,Z_q$ from *N*(0,1), and let $Z=(Z_1,\ldots,Z_q)$.
5. Let $\beta^* = \hat{\beta} + \sigma^* V^{1/2} Z$.
6. Draw n_{mis} independent variables $z_1, \ldots, z_{n_{mis}}$ from *N*(0,1), and let $e^* = \sigma^* z$, with $z = \left(z_1, \ldots, z_{n_{mis}}\right)$.
7. Let $Y^*_{mis} = X_{mis}\beta^* + e^*$.

In steps 1 to 5, the parameter values for the regression model are drawn from its posterior distribution given the observed data using non-informative priors. For reference see Appendix F, [1] and [2].

In this way, the extra uncertainty due to the fact that the regression parameters can be estimated, but not determined, from Y_{obs} and X_{obs} is reflected. Using estimated regression parameters rather than those drawn from its posterior distribution results in improper imputation, in the sense that the between imputation variance is under-estimated. In steps 6 and 7, the parameters drawn from its posterior distribution are used together with the covariates X_{mis} to generate the imputation Y^*_{mis}.

Incompletely observed covariates

Let *y* be an imputation variable, and let $x_1, \ldots, x_p$ be the incompletely observed covariates for *y*. Let R_j be the response indicator for x_j.

The variable R_j is defined by:

$$R_j = \begin{cases} 1 & \text{if } x_j \text{ is observed} \\ 0 & \text{if } x_j \text{ is not observed} \end{cases}$$

The indicator method is based on the following statistical model for *y*:

$$y = \beta_0 + \beta_{01}(1 - R_1) + \cdots + \beta_{0p}(1 - R_p) + \beta_1 R_1 x_1 + \cdots + \beta_p R_p x_p + \varepsilon \text{ ; with } \varepsilon \sim N(0, \sigma^2)$$

In this model, the term:

$\beta_j R_j x_j$ is zero when x_j is missing and is equal to $\beta_j x_j$ when x_j is observed.

When x_j is missing, the intercept term is adjusted by the term:

$\beta_{0j} (1 - R_j)$.

If a covariate x_j is completely observed, then the corresponding term $\beta_{0j} (1 - R_j)$ disappears.
By adjusting the data matrix *X*, the algorithm shown in "Completely Observed Covariates" can be applied.

Let *c* be the number of incompletely observed covariates and *i(1),…,i(c)* be the index number of these covariates. Let *X* be the adjusted data matrix constructed as follows:

1. The first column of *X* consists of 1's;
2. The j+1-th column of *X*, with $1 \leq j \leq c$, consists of 1's and 0's such that the v-th entry of this column equals 0 when the v-th data entry of $x_{i(j)}$ is observed, and is equal to 1 when this entry is missing;
3. For the *c*+1+j-th column of *X*, the i-th entry is equal to the i-th entry of x_j when this entry is observed, and is equal to 0 when this entry is missing.

Let Y_{obs} and Y_{mis} be the observed and missing data for y respectively. Let X_{obs} and X_{mis} be the rows of *X* corresponding to Y_{obs} and Y_{mis}, respectively. Each imputation Y_{mis}* for Y_{mis} is independently generated according to the same algorithm described in "Completely Observed Covariates" above.

Appendix D: Discriminant Multiple Imputation

Discriminant Multiple Imputation

This appendix describes the method used to impute binary and categorical variables for Discriminant Multiple Imputation.

Discriminant Multiple Imputation is a model based method for binary or categorical variables. The detailed imputation method is described in the following:

Let 1,…,*s* be the categories of the categorical imputation variable *y*. By applying Bayes' Theorem, the statistical model of discriminant imputation is given by the following equation:

$$P(y = j \mid x) = \frac{\phi(x \mid \mu_j; \Sigma_j)\pi_j}{\sum_{v=1}^{s} \phi(x \mid \mu_v; \Sigma_v)\pi_v}, \; j = 1, \ldots, s$$

In this equation $P(y = j \mid x)$ is the probability that the imputation variable *y* is equal to its *j*-th category given the vector *x* of the observed values of the covariates of *y* and $\phi(. \mid \mu, \Sigma)$ is the density of the multivariate normal distribution with mean, μ and covariance matrix, Σ.

μ_j and Σ_j are the conditional mean and covariance matrix of the covariates of *y* given that *y* is equal to its *j*-th category, and π_j is the apriori probability that *y* is equal to its *j*-th category.

The imputation scheme for discriminant multiple imputation is given by:

(i) Let n_j be the number of observed values of *y* equal to the *j*-th category of *y* and let $a_j = ½ + n_j$, for $j = 1,\ldots,s$;

(ii) Draw $\theta_1^*, \ldots, \theta_s^*$ from the standard Gamma distribution with parameters given by $a_1,\ldots,a_s$;

(iii) Let $\pi_j^* = \theta_j^* \Big/ \left(\sum_{v=1}^{s} \theta_v^* \right)$, for $j = 1,\ldots,s$.

(iv) For $j = 1,\ldots,s$, draw μ_j^* from the multivariate normal distribution with mean and covariance matrix given by $\hat{\mu}_j$ and S_j / n_j , where $\hat{\mu}_j$ and S_j are the sample mean and covariance matrix of the covariates of *y* calculated from the cases where *y* is observed and equal to its *j*-th category.

(v) Let $p_{ij}^* = \dfrac{\phi(X_i^T \mid \mu_j^*; S_j)\pi_j^*}{\sum_{v=1}^{s} \phi(X_i^T \mid \mu_v^*; S_v)\pi_v^*}$, for $i = 1,\ldots,n_{\text{mis}}$ and for $j = 1,\ldots,s$.

The function ϕ is the probability density function of the multivariate normal distribution given by:

$$\phi(x \mid \mu; \Sigma) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

The index *i* refers to the *i*-th missing values of *y*, *k* is the number of covariates used for imputation variable *y*, |Σ| is the determinant of Σ, and X_i^T is the row vector of observed values for the covariates of *y* corresponding to the *i*-th missing value of *y*.

(vi) Let y_i^* equal to *j* with probability p_{ij}^* $i = 1,\ldots,n_{\text{mis}}$ and for $j = 1,\ldots,s$. This is realized by drawing *u* from the standard uniform distribution and setting y_i^* equal to *j* if:

$$\sum_{v=1}^{j-1} p_{iv}^* < u \leq \sum_{v=1}^{j} p_{iv}^*$$

(vii) Impute y_i^* for the *i*-th missing data entry of *y* for $i = 1,\ldots,n_{\text{mis}}$.

In steps (i) to (iii) the probabilities π_j^* are drawn from a Diriclet distribution, which is the posterior distribution of these probabilities with non-informative prior as described in chapter 4 of "Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data sets", Brand J.P.L.

In step (iv), the means μ_j^* are randomly drawn from its normal posterior distribution. The estimated covariance matrices S_j are used in step (iv) instead of the covariance matrices drawn from a posterior distribution. Drawing the covariance matrices from their inverted Wishart posterior distribution is relatively expensive computationally.

In predicted mean single imputation, for each missing data entry the category with the largest conditional probability given the observed values of the covariates is imputed. The imputation scheme for discriminant single imputation in case of predicted mean imputation is obtained from the imputation scheme for discriminant multiple imputation as follows:

In step (v); μ_j^* is replaced by $\hat{\mu}_j$, μ_v^* is replaced by $\hat{\mu}_v$, π_j^* is replaced by n_j / n_{obs}, π_v^* is replaced by n_{vj} / n_{obs}, and p_{ij}^* is replaced by p_{ij}, where n_{obs} is the number of observed values of the imputation variable.

Step (vi) is replaced by "Let $\hat{y}_i$ be equal to the category *j*, which maximizes the probability p_{iv} for $v = 1,\ldots, s$".

In step (vii); y_i^* is replaced by $\hat{y}_i$.

Appendix E: Propensity Score Multiple Imputation

PROPENSITY SCORE MULTIPLE IMPUTATION
DIVIDE PROPENSITY SCORE INTO C QUANTILE SUBSETS
USE C CLOSEST MATCHING CASES
USE D% CLOSEST MATCHING CASES

Propensity Score Multiple Imputation

An implicit model approach based on Propensity Scores and an Approximate Bayesian Bootstrap is used to generate the imputations. The multiple imputations are independent repetitions from a Posterior Predictive Distribution for the missing data, given the observed data.

The imputation scheme is described below:

(i) The regression coefficient *b* of the logistic regression model of the response indicator R_y of the imputation variable *y* on the selected covariates including the intercept term are estimated.

(ii) To each case, a propensity score is assigned which is equal to $X_i^T b$ with *i* the index number of this case and X_i^T a row vector with its first element equal to 1 and the other element containing the observed values of the selected covariates of the *i*-th case, is assigned.

(iii) The cases in the data set are sorted according to their propensity score in ascending order.

(iv) For each missing data entry of *y*, a subset of observed values of *y* (its donor pool) is found such that their assigned propensity scores that are close to the assigned propensity score of the missings to be imputed.
This subset of observed values can be defined in different ways depending on the selected option.
Possible options are:
- Divide propensity score into c quantile subsets
- Use c closest matching cases
- Use d% closest matching cases
- Use a refinement variable.

These options are described later in this Appendix.

(v) For each missing value of *y*, the imputations are generated from its donor pool according to the Approximate Bayesian Bootstrap Method.

The estimated probability that a value of *y* is missing from the logistic regression model is a Monotone non-increasing function of the propensity score given by:

$$P(y \text{ is missing}) = 1 - \frac{\exp(\text{propensity-score})}{1 + \exp(\text{propensity-score})}$$

This implies that if instead of assigning the propensity scores to the cases, the estimated probabilities that *y* is missing are assigned to the cases. The resulting imputation method is equivalent to the one described above. That the propensity scores are used rather than these estimated probabilities is for reasons of numerical stability.

Divide propensity scores into c Quantile subsets

Using the options in the Donor Pool window, the cases of the data sets can be subdivided into *c* subsets according to the quantiles of the assigned propensity scores, where *c*=5 is the default value of *c*. This is done by sorting the cases of the data sets according to their assigned propensity scores in ascending order, as shown by the following:

The *i*-th sub-set will consist of the cases from the $\left[\frac{n}{c}*(i-1)+1/2\right]$+1-th case until the $\left[\frac{n}{c}*i+1/2\right]$-th case in the sorted data set for $i = 1,\ldots,c$, where $[x]$ is the integer part of x.
For each missing data entry of *y*, the set of observed values of *y* used to generate the imputations are the observed values of the sub-set of cases where this missing data entry belongs.

Use c Closest Matching Cases

There are two approaches to finding the *c* closest matching cases. For each missing data entry, $y_{mis}^{(i)}$, where the index *i* refers to the *i*-th missing data entry of *y*. The subset of observed values used for generating the imputations for the missing entry are the [*c*/2] observed values before, and the [*c*/2+1/2] observed values of *y*, after the missing value to be imputed (after sorting on propensity). The initial values of *y* are the observed values with an assigned propensity score closest to, and lower than, the propensity score assigned to $y_{mis}^{(i)}$. Then the [*c*/2+1/2] observed values of *y* after $y_{mis}^{(i)}$ are the observed values of *y* with an assigned propensity score closest to, and higher than the propensity score assigned to the missing data entry.

If less than [*c*/2] observed values have an assigned propensity score smaller than the assigned propensity score, then only these values are used as the observed values of *y* in the imputation. Similarly, if less than [*c*/2+1/2] observed values of *y* have an assigned propensity score larger than the assigned propensity score, then only these values are used as the observed values of *y* in the imputation.

Alternatively the difference between propensity scores will be calculated and the *c* cases with the smallest difference will be used as the donor pool. This method involves more calculations and will be computationally more intensive. With very large mounts of data it may prove more efficient to use the method described above.

Use d% Closest Matching Cases

The same as for "c Closest Matching Cases", where *c* is equal to $[(d/100)*n_{obs}]$, and where n_{obs} is equal to the number of observed values of *y*. There must be at least two values in each sub-group.

Appendix F: Mahalanobis Distance Multiple Imputation

MAHALANOBIS DISTANCE MULTIPLE IMPUTATION
CENTRE VARIABLE
USE C CLOSEST MATCHING CASES
USE D% CLOSEST MATCHING CASES

Mahalanobis Distance Multiple Imputation

For each case containing a missing value the Mahalanobis Distance D_M between that case and all other cases within the dataset, (or group, if a grouping variable has been used) is calculated. The distance is calculated using covariates specified where y is the vector of the covariates for the case with the missing value and x_i is the vector for the i^{th} fully observed case in the dataset.

$$D_M(x_i, y) = \sqrt{(x_i - y)^T S^{-1}(x_i - y)}$$

S is the covariance matrix for the set of covariates being used in the calculation of the Mahalanobis distance.

Centre Variable

When calculating the Mahalanobis Distance, the covariance matrix used is a weighted average taken across the different levels of the Centre Variable. For instance, assume we are calculating the MD between two cases (x and y). If there are three levels of the Centre Variable then when calculating the Mahalanobis Distance, D_M the following Covariance matrix (S) would be used:

$$S = \frac{(A(a-1)) + (B(b-1)) + (C(c-1))}{(a+b+c) - 3}$$

A, B and C are the three covariance matrices from within each of the three levels of the Centre Variable
a, b and c are the numbers of cases within each level of the Centre Variable

Use c Closest Matching Cases

Once the Mahalanobis Distances have been calculated the c cases that have the shortest distance from the case to be imputed are used as the donor pool.

Use d% Closest Matching Cases

The same as for "c Closest Matching Cases", where c is equal to $[(d/100)*n_{obs}]$, and where n_{obs} is equal to the number of observed values of y. There must be at least two values in each sub-group.

Appendix G: References

MULTIPLE IMPUTATION REFERENCES

SOLAS™ References

[1] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

[2] Gelman, A., Carlin, J., Stern, H., and Rubin, D.B. (1995). Bayesian Data Analysis. New York: Chapman and Hall

[3] Rubin, D.B., and Schenker, N. (1991).Multiple Imputation in Health-Care Data Bases: An Overview and Some Applications. *Statistics in Medicine* **10**, 585-598.

[4] Lavori, P., Dawson. R., and Shera, D. (1995). A Multiple Imputation Strategy for Clinical Trials With Truncation of Patient Data. *Statistics in Medicine* **14**, 1913-1925.

[5] Rubin, D.B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association* 91, 473-489.

[6] Anderson, T.W. (1957). Maximum likelihood estimates for the mulitvariate normal distribution when some observations are missing. JASA 52, 200-203.

[7] Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. Journal of the American Statistical Association 69, 467-474.

[8] Little, R.J.A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business and Economic Statistics* 6, 287-301.

[9] Mahalanobis, P C (1936). "On the generalised distance in statistics". *Proceedings of the National Institute of Sciences of India* 2 (1): 49–55.

Multiple Imputation and Related Literature References

Box, G.E.P., Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Reading, Mass: Adisson Wesley.

Chand, N. and Alexander, C.H. (1994). *Imputing Income for An N-Person Consumer Unit*, Bureau of the Census paper presented at the American Statistical Association Annual Meeting in Toronto.

Clogg, C., Rubin. D.B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple Imputation of Industry and Occupation Codes in Census Public Use Samples Using Bayesian

Logistic Regression, *Journal of the American Statistical Association* **86**, 68-78.

Efron, B. (1994). Missing Data, Imputation, and the Bootstrap (with discussion*). Journal of the American Statistical Association* **89**, 463-478.

Efron, B., and Tibsharani, R. (1993). Assessment of Reported Differences Between Expenditures and Low Incomes in the U.S. Consumer Expenditure Survey. Paper presented at the American Statistical Association Annual Meeting, Toronto.

Ezzati-Rice, T.M, Johnson, W., Khare, M., Little, R.J.A., Rubin, D.B., and Schafer, J.L. (1995). A Simulation Study to Evaluate The Performance of Multiple Imputations in NCHS Health Examination Survey, in Proceedings of the Bureau of the Census Eleventh Annual Research Conference, pp. 257-266.

Ezzati-Rice, T.M., Khare, M., and Schafer, J.L. (1993). Multiple Imputation of Missing Data in NHANES III, paper presented at the American Statistical Association Annual Meeting, San Francisco.

Fahimi, M., and Judkins, D. (1993). Serial Imputation of NHANES III With Mixed Regression and Hot-Deck Technique, paper presented at the American Statistical Association Annual Meeting, San Francisco.

Fay, R.E. (1990). VPLX: Variance Estimation for Complex Surveys: Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 266-271.

Fay, R.E. (1991). A Design-Based Perspective on Missing Data Variance, in Proceedings of the1991 Annual Research Conference, U.S. Bureau of the Census, pp. 429-440.

Fay, R.E. (1992). When are Inferences from Multiple Imputation Valid?, in Proceedings of the Survey Research Methods Sections, American Statistical Association, pp. 227-232.

Fay, R.E. (1993). Valid Inferences from Imputed Survey Data, paper presented at the Annual Meeting of the American Statistical Association, San Francisco.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. *Journal of the American Statistical Association* this issue, 490-498.

Fisher, R.A. (1925). Theory of Statistical Estimation, Proceedings of the Cambridge Philosophical Society, 22. 700-725.

Fisher, R.A. (1934). Discussion of on the Two Different Aspects of the Representative Method of Stratified Sampling and the Method of Purposive Selection, by J. Neyman. *Journal of the Royal Statistical Society* **Ser. A. 97**, 614-619.

Freedman, V. (1990). Using SAS to Perform Multiple Imputation: Discussion Paper Series UIPSC-6, The Urban Institute, Washington, DC.

Gelfand. A.E., and Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating MarginalDensities. *Journal of the American Statistical Association* **85**, 398-409.

Gelman, A. and Rubin, D.B. (1992). Inference from Iterative Simulation Using MultipleSequences (with discussion) *Statistical Science* **7**, 457-472.

Hansen, M.H. (1987). A Conversation with Morriss Hansen (I. Olkin, interviewer). *Statistical Science* **2**,162-179.

Heitjan, D.F., and Rubin, D.B., (1990). Inference from Coarse Data via Multiple Imputation with Application to Age Heaping. *Journal of the American Statistical Association* **85**, 304-314.

Herzog, T., and Lancaster, C. (1980). Multiple Imputation Modeling for Individual Social Security Benefit Amounts, in Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 398-403.

Huber, P.J. (1976). The Behavior Maximum Likelihood Estimates Under Non-standard Conditions, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press, pp. 221-233.

James, I.R. (1995). A Note on the Analysis of Censored Regression Data by Multiple Imputation. *Biometrics* **51**, 358-362.

Johnson, C.L., Curtin, L.R., Ezzati-Rice, T.M., Khare, M., and Murphy, R.S. (1993). Single and Multiple Imputation: The NHANES Perspective, paper presented at the Annual Meeting of the American Statistical Association, San Francisco.

Kalton, G. (1983). Compensating for Missing Survey Data, Ann Arbor, MI: Institute of Social Research, University of Michigan.

Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation, in Proceedings of the Survey Research Methods Section of the American Statistical Association, pp.1-10.

Kong, A., Liu, J., and Wong, W.H. (1994). Sequential Imputation and Bayesian Missing Data Problems. *Journal of the American Statistical Association* **89**, 278-288.

Kott, P.S. (1992). A Note on a Counter-Example to Variance Estimation Using Multiple Imputation, technical report, U.S. National Agriculture Service.

Krewski, D., and Rao, J.N.K. (1981). Inference from Stratified Samples: Properties of the Linearisation, Jackknife, and Balanced Repeated Replication Methods. *The Annals of Statistics* **9**, 1010-1019.

Lehmann, E.L. (1959). Testing Statistical Hypotheses, New York: John Wiley.

Li, K.H, Meng, X.L., Raghunathan, T.E., and Rubin, D.B. (1991). Significance Levels from Repeated p-Values With Multiple-Imputed Data. *Statistica Sinica* **1**, 65-92.

Li, K.H, Raghunathan, T.E, and Rubin, D.B. (1991). Large Sample Significance Levels from Multiply-Imputed Data Using Moment-Based Statistics and an F Reference Distribution. *Journal of the American Statistical Association* **86**, 1065-1073.

Little, R.J.A. (1979). Maximum Likelihood for Multiple Regression With Missing Values: A Simulation Study. *Journal of the Royal Statistical Society* **B41**, 76-87.

Little, R.J.A. (1988). Missing Data in Large Surveys (also with discussion). *Journal of Business and Economic Statistics* **6**, 287-301.

Little, R.J.A., and Rubin, D.B. (1987). Statistical Analysis with Missing Data, New York: John Wiley.

Little, R.J.A., and Rubin, D.B. (1993). Assessment of Trial Imputations for NHANES III, project report, Datametrics Research, Inc.

Liu, C. and Rubin, D.B. (1996). M: Multiple Imputation System, report, Datametrics Research, Inc.

Liu, J.S., and Chen, R. (1995). Blind De-convolution via Sequential Imputations. *Journal of the American Statistical Association* **90**, 567-576.

Meng, X. (1994). Multiple Imputation with Uncongenial Sources of Input (with discussions). *Statistical Science* **9**, 538-574.

Meng, X.L., and Rubin, D.B. (1992). Performing Likelihood Ratio Tests with Multiply Imputed Data sets. *Biometrika* **79**, 103-111.

Miller, R.G. (1974). The Jackknife - A Review. *Biometrika* **61**, 1-17.

Mislevy, R.J., Johnson, E.G., and Muraki, E. (1992). Scaling Procedures in NAEP. *Journal of Educational Statistics* **17**,131-154.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society* **A97**, 558-606.

Paulin, G.D., and Ferraro, D.L. (1994). Do Expenditures Explain Income? A Study of Variables for Income Imputation, paper presented at the Annual Meeting of the American Statistical Association, Toronto.

Rao, J.N.K., and Shao, J. (1992). Jackknife Variance Estimation With Survey Data Under Hot Deck Imputation. *Biometrika* **79**, 811-822.

Rubin, D.B. (1977a). Formalizing Subjective Notions about the Effect of Non-respondents in Sample Surveys. *Journal of the American Statistical Association* **72**, 538-543.

Rubin, D.B. (1977b). The Design of a General and Flexible System for Handling Non-Response in Sample Surveys, working document prepared for the U.S. Social Security Administration.

Rubin, D.B. (1978). Multiple Imputations in Sample Surveys-A Phenomenological Bayesian Approach to Non-response, in Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 20-34. (See also Imputation and Editing of Faulty or Missing Survey Data, US Department of Commerce, pp. 1-23).

Rubin, D.B. (1980). Handling Non-response in Sample Surveys by Multiple Imputations. Monograph, U.S. Department of Commerce, Bureau of the Census.

Rubin, D.B. (1981). The Bayesian Bootstrap. *The Annals of Statistics* **9**, 130-134.

Rubin, D.B. (1983). Progress Report on Project For Multiple Imputation of 1980 Codes, manuscript distributed to the U.S. Bureau of the Census, the U.S. National Science Foundation, and the Social Science Research Foundation.

Rubin, D.B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics* **12**, 1151-1172.

Rubin, D.B. (1988). Using the SIR Algorithm to Simulate Posterior Distributions (with discussion), in Bayesian Statistics 3, eds. J.M. Bernard, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, New York: Oxford University Press, pp. 395-402.

Rubin, D.B. (1990). Imputation Procedures and Inferential Versus Evaluative Statistical Statements, in Proceedings U.S. Census Bureau Sixth Annual Research Conference, pp. 676-679.

Rubin, D.B., and Schenker, N. (1991). Analyzing Multiple Imputed Data sets.

Rubin, D.B. (1993). Satisfying Confidentiality Constraints Through the Use of Synthetic Multiple-Imputed Micro-Data. *Journal of Official Statistics* **9**, 461-468.

Rubin, D.B. (1994). Comments on Missing Data, Imputation, and the Bootstrap by B. Efron: *Journal of the American Statistical Association* **89**, 485-8.

Rubin, D.B., and Schenker N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples With Ignorable Non-response. *Journal of the American Statistical Association* **81**, 366-374.

Rubin, D.B., and Schenker N. (1987). Interval Estimation from Multiply Imputed Data: A Case Study using Agriculture Industry Codes. *Journal of Official Statistics* **3**, 375-387.

Schafer, J.L. (1996). Analysis of Incomplete Multivariate Data by Simulation, New York: Chapman and Hall.

Schafer, J.L., and Schenker, N. (1991). Variance Estimation with Imputed Means: Proceedings of the Survey Research Methods Section, American Statistical Association pp. 696-701.

Schafer, J.L., Kare, M., Little, F.J.A., and Rubin, D.B. (1993). Multiple Imputation of NHANES III, paper presented at the Annual Meeting of the American Statistical Association, San Francisco.

Schenker, N. (1989). The Use of Imputed Probabilities for Missing Binary Data, in Proceedings of the 5th Annual Research Conference, Bureau of the Census, pp. 133-139.

Schenker, N., Treiman, D.J., and Weidman, L. (1993). Analyses of Public Use Decennial Census Data with Multiply Imputed Industry and Occupation Codes. *Applied Statistics* **42**, 545-556.

Smith, A.F.M., and Gelfand, A.E. (1992). Bayesian Statistics Without Tears: A Sampling- Resampling Perspective. *The American Statistician* **46**, 84-88.

Tanner, M.A., and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528-550.

Treiman, D.J., Bielby, W., and Cheng, M. (1987). Significance Levels from Public Use Data With Multiply-Imputed Industry Codes, unpublished doctoral thesis, Harvard University, Dept. of Statistics.

Treiman, D.J., Bielby, W., and Cheng, M. (1989). Evaluation of a Multiple Imputation Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard, Sociological Methodology, 18, 309-345.van Buuren, S., van Mulligen, E.M.,, and Brand, J.P.L.

Treiman, D.J., Bielby, W., and Cheng, M. (1995). Omgaan Met Ontbrekende Gevevens in Statistische Databases: Multiple Imputatie in HERMES, Kwantitatieve Methadone, 50, 503-504.

van Buuren, S., van Rijckevorsel, J.L.A., Rubin, D.B., Treiman, D.J., Bielby, W., and Cheng, M. (1993). Multiple Imputation by Splines, in Bulletin of the International Statistical Institute, Contributed Papers II, 503-504. Weld, L. Wolter, K.M. (1984). Introduction to Variance Estimation. New York: Springer-Verlag.