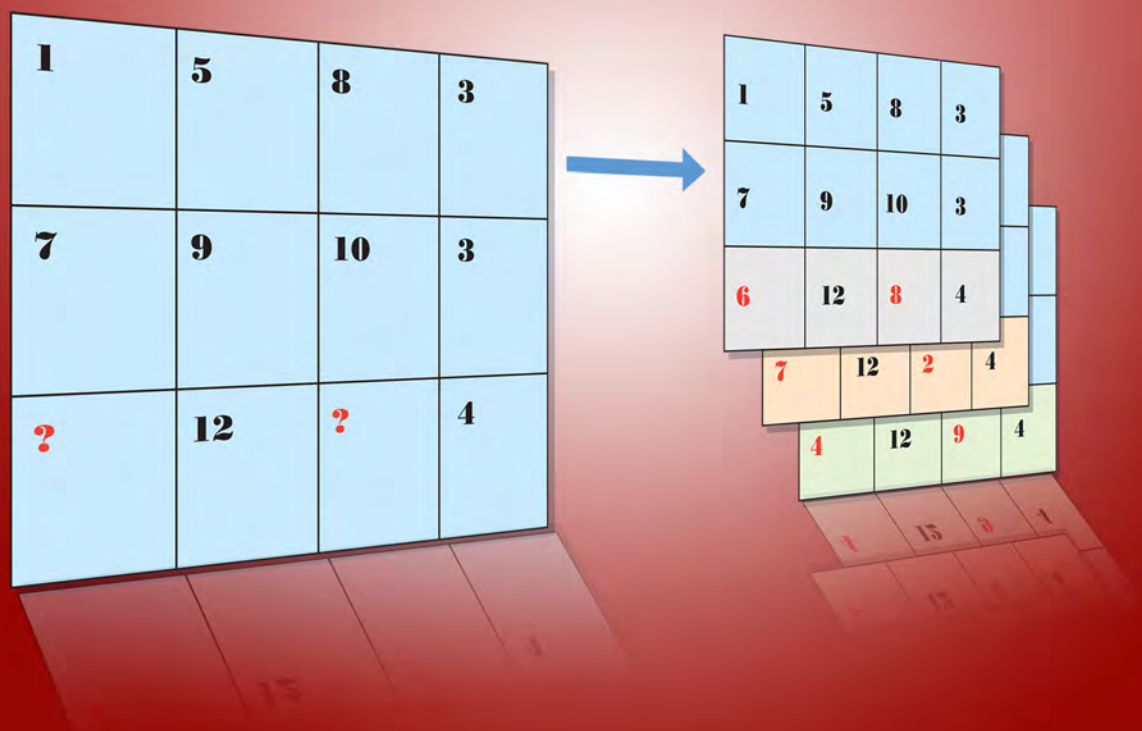
# Multiple Imputation of Missing Data Using SAS®




Patricia Berglund and Steven Heeringa



# Contents



# Chapter 1: Introduction to Missing Data and Methods for Analyzing Data with Missing Values



## 1.1 Introduction

Over the past half-century, statistical analysts have employed a wide range of techniques to address the theoretical and practical question of "what do I do about missing values?" These techniques have ranged from a simple default of dropping observations with missing data values from analysis (the list-wise deletion default in SAS and most other major software systems) to highly sophisticated methods for modeling the missing data mechanism in order to derive imputed values or to conduct a complex maximum likelihood analysis. There is no single approach that is optimal for all missing data problems—either in theory or in practice. Fortunately for the data analyst, SAS and other major statistical analysis software packages now provide their users with robust procedures tailored to address differing problems of missing data. *Multiple Imputation of Missing Data Using SAS* is written to serve as a practical guide for those dealing with general missing data problems in fields such as the social, biological, and physical sciences; medical and public health research; education; business; and many other scientific and professional disciplines. Central to this book is the method of multiple imputation (MI) for item missing data. Supported by the SAS PROC MI and PROC MIANALYZE procedures, MI is based on a powerful set of methods for both filling in the missing values in the user data set and for performing robust statistical estimation and inference using the completed data sets.

The combination of basic theoretical background and extensive practical applications using SAS (v9.4) presented in this volume provides a solid foundation for understanding and resolving missing data problems using the multiple imputation method. The applications presented in Chapters 4 through 8 address a number of common missing data problems and imputation approaches using PROC MI and PROC MIANALYZE along with various descriptive and inferential tools for analysis of complete data sets. All examples stress the three-step process of multiple imputation: 1) selection of an appropriate data model for the imputation and application of an appropriate imputation method using PROC MI; 2) analysis of complete data sets using standard or SURVEY procedures; and 3) synthesis of analytic results for statistical inference using PROC MIANALYZE.

## 1.2 Sources and Patterns of Item Missing Data

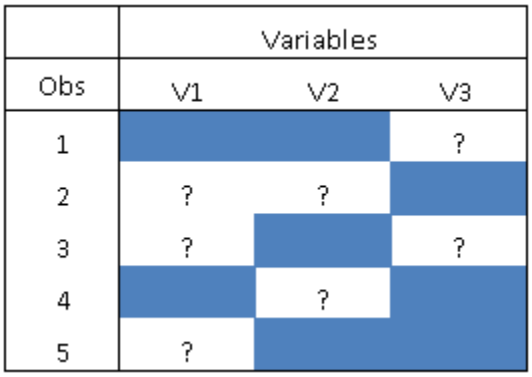
Missing data takes many forms and can be attributed to many causes. For example, in data derived from surveys, item missing data occurs when a respondent elects not to answer certain questions, resulting in only a "don't know" or "refused" response. The failure to respond can be driven by the desire to elude answering questions that are sensitive in nature or perceived to be intrusive (e.g., illegal behavior, income-related, or medical history questions) or by a simple lack of knowledge required to answer the questions (e.g., expenditure on durable goods in the past 30 days). More generally in observational and experimental research, missing observations can arise due to missed clinical appointments, equipment failures, or other circumstances that disrupt the intended measurement. For example, a power outage that deactivates environmental data collection instrumentation for a period of time results in a missing data problem. In some research settings such as epidemiology (Schenker et al. 2010) and genetics (Bobb et al. 2011), "missing data" is actually better labeled "nonobserved data" and powerful new statistical tools including MI are used to impute the nonobserved data based on strong associations with variables that were observed in the same or a different study.

Missing data problems can be classified based on a notation and taxonomy employed by Little and Rubin (2002). In regard to the notation, the data of analytic interest are the true underlying values for an $n$ x $p$ matrix, $Y$, consisting of $i$=1,…,$n$ rows (sample cases) and $j$=1,… ,$p$ variables, $Y_i = \{Y_{i1}, \ldots, Y_{ip}\}$ for each case. This underlying set of true values for the variables of interest is then decomposed into two subvectors of variables, $Y = \{Y_{obs}, Y_{mis}\}$, where $Y_{obs}$ are the set of variable values that are observed and $Y_{mis}$ are the variable values that are missing and must be imputed (Heeringa, West, and Berglund 2010). The statistical properties of individual variables and their relationships in the underlying data are governed by a distributional model, $f(y|\theta)$. Our analytic interest lies in making inference about the parameters $\theta$ or functions of these parameters. Maximum likelihood (ML) methods are often used in analysis, although corresponding Bayesian methods of analysis and inference may also be employed. In addition to the distributional model for the underlying data, the missing data problem also includes a second statistical model, $g(m|\psi)$, that governs the stochastic process which determines whether the true value of a $Y_{ij}$ is missing ($M_{ij}$=1) or observed ($M_{ij}$=0). Corresponding to the $Y$ matrix of true data values is a second array of identical dimension, $M$, of indicator values that flag whether the corresponding element of $Y$ is missing or observed.

To help us better understand our missing data problem and to choose an appropriate imputation strategy, statisticians have created a two-part taxonomy for the most common missing data problems. The first dimension is labeled the *missing data pattern*, which as the name implies defines the particular distribution of the missing observations (represented by the $M$ matrix) across the data cases ($i$=1,…,$n$) and variables ($j$=1,…,$p$) that comprise our analytical data set.

The most common missing data pattern is termed generalized or arbitrary—where there is no particular pattern in the missing data structure. As illustrated by the cells with "?" in the schematic array in Figure 1.1, missing observations are distributed across cases and variables in a nonsystematic fashion. The arbitrary missing data pattern illustrated in Figure 1.1 might be handled with an imputation using a Markov chain Monte Carlo (MCMC) or a fully conditional specification (FCS) method, both available in PROC MI.

**Figure 1.1: Generalized Pattern of Missing Data**



Some data collection processes produce a more structured or systematic pattern of missingness in the data. For example, in a medical study with multiple phases, missing data can occur when an entire phase of the data collection effort, such as a blood draw or obtaining medical records needed for follow-up, requires special consent of the subject. Without agreement from the study subject for participation in this type of data collection, there is missing data for all variables from the entire phase of the study. The result is a monotonic missing data pattern similar to that illustrated in Figure 1.2. Nonresponse to follow-up waves in a longitudinal survey may also produce a monotonic pattern of item missing data. Monotone patterns of missing data lend themselves to imputation methods that require simpler assumptions than approaches that are required for a general pattern of missing data and are efficiently handled in PROC MI. In fact, in Chapter 6 we will consider a two-step procedure that imputes small amounts of missing data in selected variables in order to transform a problem with a generalized missing data pattern to one that has a monotone pattern for key variables that suffer the highest rates of missing data. We note here that analysts always have the option to address monotone missing data patterns using procedures such as the FCS method that are applicable to the more general case of an arbitrary missing data pattern.

**Figure 1.2: Monotone Missing Data Pattern**

| | Variables | | |
|---|---|---|---|
| Obs | V1 | V2 | V3 |
| 1 | | | |
| 2 | | | |
| 3 | | | ? |
| 4 | | ? | ? |
| 5 | ? | ? | ? |

A third pattern of missing data arises in studies that incorporate randomization procedures to allow item missing data on selected variables for subsets of study observations. The technique is termed *matrix sampling* or "missing by design" sampling (Thomas et al. 2006) and is often employed with modularized sets of data observations (e.g., physical tests or sets of survey questions). In a survey data collection that employs matrix sampling, designated subsets of core questions are asked of all respondents, with additional modules of more in-depth questions randomly assigned to subsamples of participants. Matrix sampling designs tend to produce a non-monotonic missing data structure such as that illustrated in Figure 1.3.

**Figure 1.3 Matrix Sampling (Missing by Design)**

| | Variables | | |
|---|---|---|---|
| Obs | V1 | V2 | V3 |
| 1 | | ? | |
| 2 | | ? | |
| 3 | | | ? |
| 4 | | | ? |
| 5 | | | ? |

For this type of missing data pattern, multiple imputation has typically been the primary tool to analyze these data (Raghunathan and Grizzle 1995). Use of PROC MI with an imputation method such as FCS that is suitable

for an arbitrary missing data pattern can be employed. As with any imputation problem, the recommended imputation method depends on the pattern of missing data and the type of variables to be imputed.

## 1.3 Item Missing Data Mechanisms

The second dimension of statisticians' two-part taxonomy of missing data problems is termed the *missing data mechanism*. Missing data for a single variable is classified into one of three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Missing data are MCAR if the probability that an item value is missing is completely random and does not depend on the missing values for a case, $Y_{mis}$, nor does it depend on any of the observed variables for the case, $Y_{obs}$. A more realistic assumption for item missing data that underlies the procedures covered in this book is that the data are missing at random (MAR). The MAR assumption requires that, conditional on the observed data for the case, $Y_{obs}$, the probability that a value is missing does not depend on the true values of the missing items, $Y_{mis}$. For example, the predictive distribution used to draw imputed values for $Y_{mis}$ may be a regression model in which the predictors are selected from $Y_{obs}$. In practice, the MAR assumption may not strictly apply for all missing items. If the probability that a variable value is missing depends on the missing value and cannot be fully explained by the remaining observed variables, $Y_{obs}$, the missing data mechanism is labeled missing not at random (MNAR). For example, if after accounting for observed age, gender, education, and marital status of the household head, the probability of item missing data on a measure of household income depends on the underlying dollar value, then the problem is MNAR.

Little and Rubin (2002) classify a missing data mechanism as "ignorable" for likelihood-based inference if two conditions hold: 1) the missing data are MAR (missingness does not depend on $Y_{mis}$); and 2) the parameters of the data distribution, $f(y|\theta)$, are distinct from the parameters of the model for the missing data mechanism $g(M|\psi)$. (For Bayesian inference, the second condition requires that the prior distributions for $\theta$ and $\psi$ are independent.) From our perspective as analysts of data with missing values, the first of these two conditions is the more important. Likelihood inference remains valid, albeit statistically less efficient, if the parameter spaces of the data distribution and the missing data generating model are not distinct.

A commonly asked question is, "Can the MAR assumption be tested in SAS?" The current version of SAS/STAT software, SAS/STAT 13.1, implements sensitivity analysis for departures from the MAR assumption with the new MNAR statement in PROC MI. See the PROC MI documentation for details. In addition, Little (1988) presents a likelihood ratio test of the null hypothesis that the data are MCAR versus the alternative that they are MAR (conditional on a defined set of observed covariates). Although useful in some special cases, this test is not sufficient to establish that the missing data mechanism is ignorable. The econometric literature on selection bias (e.g., Amemiya, 1985; Heckman, 1976) presents several tests of the null hypothesis that the data are MAR versus the MNAR alternative. As Little (1985) notes, however, these tests are very sensitive to correct model specification. Schafer (1997) describes a number of data collection designs where the missing data mechanism is clearly ignorable (double sampling, medical screening with multiple tests, matrix sampling, etc.) and others where it is not clear whether the missing data mechanism can be ignored (sample surveys where people are not at home, unexpected problems in experiments that prevent data collection, etc.). One approach to address possible departures from the MAR assumption is to increase the number of variables in the imputation model, thus making the assumption more plausible (Schafer 1997). Extreme departures from the MAR assumption may require special methods that require explicit modeling of the missing data mechanism. Imputation methods for MNAR problems are beyond the scope of this volume, and therefore we refer the reader to the statistical literature on potential methods. Chapter 15 of Little and Rubin (2002) is a good starting point to learn more about methods for MNAR missing data.

## 1.4 Review of Strategies to Address Item Missing Data

### 1.4.1 Complete Case Analysis

The first and simplest approach to the problem of missing data is complete case analysis. The majority of SAS analytic procedures default to list-wise deletion and will automatically strike any case with item missing data from the analysis. Many analysts faced with a missing data problem question if imputation (or an alternative missing data procedure) is really necessary when considering the additional effort required to conduct the

analysis. While item missing data rates of 1% to 5% for single variables are not likely to produce major biases for univariate estimates based on only the complete cases, list-wise deletion of missing data cases from a complex multivariate analysis can result in a significant loss of statistical information. The analytic implications of item missing data are both practical and statistical. A practical consequence of list-wise deletion is that it permits the size and composition of the analysis sample to vary depending on the variables included in the analysis, making it difficult to maintain a standardized set of inputs across a variety of analytic methods. Other statistical implications of list-wise deletion of cases due to item missing data include reduction of effective or "working" sample size and loss of precision, regardless of the missing data mechanism. If the missing data are MAR, list-wise deletion of cases can result in a biased analysis for the complete cases. Of course, if the aggregate rate of missing data or the individual rate for a single key variable (e.g., household income, diastolic blood pressure) is higher (say, 5% to 10%), the precision losses and instability in case counts for differing analyses and the potential for bias under MAR will increase. In these cases, it is best practice to employ maximum likelihood estimation methods or missing data imputation that maximize the use of the observed data in the complete and incomplete cases.

### 1.4.2 Complete Case Analysis with Weighting Adjustments

A second option for handling missing data is to analyze complete cases but introduce additional weighting adjustments to compensate for item missing data on a key variable or set of variables. The use of weighting to compensate for missing data is generally limited to monotonic missing data patterns in which large numbers of variables are missing for each case. Unit nonresponse in surveys or other data collections, phase nonresponse, and longitudinal attrition in panel surveys each produce missing data patterns where a global weighting adjustment—applicable across many forms of analysis of the data—is an appropriate and practical choice to compensate for missing observations. In practice, adjustments to the base weight variables to address item missing data on single variables often lead to difficulties, in that different adjustment factors would be needed for each target variable. Analytically, the weight-by-variable approach would work for univariate analyses, but which of the variable-specific weights would be chosen for a multivariate analysis? Another problem is that access to data used for base weight construction is restricted, making later weight adjustments impossible. For these reasons, imputation is generally considered a better strategy for addressing generalized patterns of item missing data.

### 1.4.3 Full Information Maximum Likelihood

For some item missing data problems, the preferred approach may be one of several maximum likelihood approaches that are designed to find the parameter estimates of interest, $\hat{\theta}$, that maximize the likelihood, $L(\theta|Y_{obs})$. Full information maximum likelihood (FIML) methods (Enders 2001) directly maximize a likelihood function in which incomplete cases contribute support only to estimation of parameters for which the sufficient statistics are functions of the observed values for the case (e.g., $\mu_1,\sigma_1$ for observed $Y_1$; $\sigma_{12}$ for observed $\{Y_1, Y_2\}$). In the statistical literature, many common applications of FIML are used in analyses such as structural equation modeling (SEM) or other latent variable model analyses in which the full data $Y$ are assumed to follow a multivariate normal distribution with parameter vector $\theta=\{\mu,\Sigma\}$. The FIML method is available in PROC CALIS—a standard SAS procedure for conducting SEM or other related forms of latent variable modeling. FIML methods require that the user define the parametric likelihood for the data. This presents a challenge in FIML applications to complex sample survey data where weighting is required and informative stratification and clustering of observational units make it difficult to specify the true form of the data likelihood (Heeringa, West, and Berglund 2010). We should note here that these same design features also pose challenges in multiple imputation approaches to missing data from complex sample designs (see Chapter 4).

### 1.4.4 Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm (Little and Rubin 2002) is another tool that can be used in missing data problems to generate ML estimates. Like FIML methods, the EM method requires a parametric likelihood function for the complete data. EM employs an iterative two-step (expectation and maximization) approach to numerically derive ML estimates of parameters. The E step of the algorithm replaces the missing values with their expectations under the current iteration's estimates of the distributional model and constructs the corresponding sufficient statistics/complete data log-likelihood function. The M step then maximizes the complete data likelihood to obtain updated estimates of the model parameters. This E and M cycle repeats until the model parameters converge. If the complete data for the analysis problem are assumed to be distributed as

multivariate normal, MVN($\mu$,$\Sigma$), the EM statement in PROC MI can be used to compute maximum likelihood estimates of the mean and covariance parameters. The estimated mean vector and covariance matrix can be output to a user designated file with the OUTEM= option of the EM statement in PROC MI. A "completed" data set with missing values replaced by the EM estimates of their expected values can also be output. However, it is important to note that the "imputed" data set so generated does not account for two sources of variability: 1) variability of the true values about their expectations (residual variance) and 2) the imputation variability that is inherent in the estimation of the expected values. These two sources of variability are reflected in a proper MI treatment of the missing data problem. EM does play an important role in PROC MI in that the completed data set generated by EM serves as the matrix of starting values for the iterative MCMC multiple imputation procedure.

### 1.4.5 Single Imputation of Missing Values

A common approach for handling item missing data prior to analysis is to perform a single imputation of missing values, creating a "complete" data set. In fact, scientific public use data sets are often released with a single imputed value replacing missing data on key variables. Or, data users may also choose to perform their own single imputations using an established stochastic imputation method such as the hot deck, regression imputation, or predictive mean matching (Little and Rubin 2002). Use of procedures such as mean, median, or modal value imputation are not encouraged unless the imputation is simply serving to fill in a small handful of missing values for an otherwise nearly complete variable.

An advantage to a singly imputed data set is that it is "completed" with missing values replaced by imputed values. Provided that the imputation technique is multivariate and retains the stochastic properties in the observed data, a single imputation may address potential bias for MAR missing data. On the other hand, an important disadvantage in the standard analysis of a singly imputed data set is that it precludes estimation and inference that fully reflects the variance attributable to the item missing data imputations. Rao and Shao (1992) have proposed a technique for estimating variances and developing confidence intervals for estimates based on singly imputed data sets. More recently, Kim (2011) has proposed the method of fractional imputation, which also enables variance estimation and inference from imputed data. At this writing, neither of these methods has been implemented in SAS procedures.

### 1.4.6 Multiple Imputation

The robust, flexible option in many practical problems and the major focus of this book is to address missing values within the MI framework for estimation and inference. This approach consists of a three-step process: 1) formulation of the imputation model and imputation of missing data using PROC MI, 2) analysis of complete data sets using standard SAS procedures (that assume the data are identically and independently distributed or from a simple random sample) or SURVEY procedures for analysis of data from a complex sample design, and 3) analysis of the output from the two previous steps using PROC MIANALYZE. Many types of missing data patterns and analytic models can be handled within this framework, making it, in our opinion, the preferred option for dealing with most missing data problems.

The many approaches and options available in PROC MI reflect the historical sequence of developments in MI theory and practice. Some approaches were developed specifically for particular patterns of missing data (e.g., MONOTONE), while others assume specific joint distributions for the variables of interest (e.g., MCMC for multivariate normal data). In years past, analysts often faced practical problems where the data patterns or distributions did not conform to the exact theoretical assumptions of the available multiple imputation tools. For example, variables with few missing observations could be initially imputed to convert a generalized pattern of missing data to a monotone missing data problem. The MCMC method, designed for multivariate normal data, would be applied to a generalized pattern of missing data for a vector of categorical variables or a mixture of categorical and continuous variables. While many of these MI practices are still serviceable, it is the case that new developments in PROC MI now provide a simpler, better approach. For example, we view the newly introduced FCS method as preferred over MCMC for multiple imputation of multivariate problems that include mixtures of categorical and continuous variables. To ensure comprehensive treatment of the capabilities of SAS for MI, throughout this text we will attempt to cover all of the common approaches—old and new—that are available in PROC MI, but when multiple alternatives are available we will clearly indicate which approach we judge to be current best practice.

## 1.5 Outline of Book Chapters

Chapter 1 introduced the topic of imputing missing data and the issues that arise in analysis of data sets with missing data. It provides an overview of how and why missing data occurs along with an introduction to the use of multiple imputation to deal with missing data problems. It also presents an overview of examples to come in later chapters.

Chapter 2 offers a detailed look at imputation with an emphasis on the multiple imputation approach. It addresses how to implement multiple imputation by outlining the general framework of model specification, imputation methods, analysis of MI data sets, and MI estimation and inference. A description of general imputation algorithms and formulae for the MI and MIANALYZE procedures are presented in this chapter. Chapter 2 also includes a brief overview of PROC MI and PROC MIANALYZE along with discussion of common SAS procedures used in the MI process.

Chapter 3 outlines a general step-by-step approach for planning and conducting a multiple imputation analysis in SAS. A simple example of the MI process is presented at the end of this chapter and serves as a prelude to more complex applications described in later chapters.

Chapter 4 provides an overview of the special issues involved in multiple imputation for complex sample design data. It includes a discussion of how things change with complex sample data and provides a comparative example of a naive imputation (ignoring the design in the imputation step) contrasted with imputation that includes the complex sample design features in the imputation model.

Chapter 5 covers imputation of continuous variables. Each example in this chapter includes a brief overview of the statistical foundations of the particular imputation method; demonstration of the three-step process of multiple imputation, including use of applicable options and diagnostics in both PROC MI and PROC MIANALYZE; and interpretation of the output from each step of the process.

Chapter 6 repeats the process and steps described for Chapter 5 but focuses on imputation of classification (categorical) variables.

Chapter 7 presents two case studies typical of "real-world" missing data problems and multiple imputation options for analysis. The first example provides a comparison of a complete case analysis and MI treatment of a missing data problem based on the Health and Retirement Study (HRS). The second case study demonstrates multiple imputation of longitudinal data based on a clinical trial focused on the impact of anti-epilepsy medication on seizures.

Chapter 8 includes examples of preparation of output data sets from the analysis of complete data sets in formats readable by PROC MIANALYZE. This chapter covers details of the various types of estimates, parameter, and covariance output data sets that can be used by PROC MIANALYZE and includes examples using a variety of regression procedures.

The author page for this book includes the SAS code and SAS data sets used in the application examples in the book as well as FAQs, a bibliography, and other resources and updates for data analysts using the SAS multiple imputation procedures.

## 1.6 Overview of Analysis Examples

In Chapters 4 through 7 we include a variety of applications of multiple imputation to a wide range of data sets, imputation methods, and analysis techniques. Table 1.1 is a summary grid outlining the examples, including information about the three-step process of multiple imputation. Details are provided about each example, such as chapter/example number, data set name and characteristics, missing data pattern, type of variable(s) imputed, imputation method, type of variables used in the imputation, and analysis procedure used in the second step of the MI process.

**Table 1.1: Overview of Complete Multiple Imputation Examples from Chapters 4 through 7**

| Chapter/ Example Number | Data Set/Design Characteristics | Missing Data Pattern | Type of Variable(s) Imputed | Type of Variables Used in Imputation | Imputation Methods Used | Analytic Technique Used to Analyze Imputed Data Sets |
|---|---|---|---|---|---|---|
| 4.1 | NHANES 2009–2010/Complex Sample Design | Monotone | Continuous | Mixed | Monotone | Design-adjusted means with a subpopulation (PROC SURVEY MEANS) |
| 5.2 | Major League Baseball Players' Salaries (1992)/Standard | Arbitrary | Continuous | Continuous | MCMC | Linear regression (PROC REG) |
| 5.3 | NHANES 2009–2010/Complex Sample Design | Monotone | Continuous | Mixed | Regression PMM | Design-adjusted means analysis for a specified subpopulation (PROC SURVEYMEANS |
| 5.4 | NHANES 2009–2010/Complex Sample Design | Arbitrary | Continuous | Mixed | FCS regression | Design-adjusted linear regression (PROC SURVEYREG) |
| 6.2 | Myocardial Infarction/ Standard | Monotone | Categorical | Mixed | Logistic regression | Frequency tables (PROC FREQ) |
| 6.3 | NCS-R/Complex Survey Design | Arbitrary | Categorical | Mixed | FCS logistic regression FCS discriminant function | Design-adjusted frequency tables (PROC SURVEYFREQ) |
| 6.4 | NHANES 2009–2010/Complex Sample Design | Arbitrary | Categorical | Mixed | FCS logistic regression Multistep MCMC monotone with logistic regression | Design-adjusted logistic regression (PROC SURVEY LOGISTIC) |
| 7.1 | HRS 2006/Complex Sample Design | Arbitrary | Mixed | Mixed | FCS logistic regression, discriminant function, regression | Design-adjusted logistic regression (PROC SURVEY LOGISTIC) |
| 7.2 | Clinical Trial Seizure Data/Standard Longitudinal data | Arbitrary | Continuous | Mixed | FCS PMM | Poisson Regression with Repeated Measures (PROC GENMOD) |

Table 1.2 summarizes the application demonstrated in Chapter 8 where we highlight the process of producing output data sets from step 2 that can be easily input into PROC MIANALYZE.

**Table 1.2: Overview of Creating Output Data Sets for PROC MIANALYZE from Chapter 8**

| Chapter/ Example Number | Data Set/Design Characteristics | Missing Data Pattern | Type of Variable(s) Imputed | Type of Variables Used in Imputation | Imputation Methods Used | Analysis Procedure Used to Create Output Data Set for Use in PROC MIANALYZE |
|---|---|---|---|---|---|---|
| 8.2 and 8.3.1 | MLB Baseball Players' Salaries (1992)/Standard | Monotone | Continuous | Mixed | MCMC | Linear regression (PROC GLM) |
| 8.2 and 8.3.2 | MLB Baseball Players' Salaries (1992)/Standard | Monotone | Continuous | Mixed | MCMC | Linear regression (PROC MIXED) |
| 8.4 and 8.5 | NCS-R/Complex Sample Design | Arbitrary | Categorical | Mixed | FCS logistic regression | Design-adjusted Proportional Hazards model (PROC SURVEY PHREG) |

# Chapter 2: Introduction to Multiple Imputation Theory and Methods



## 2.1 The Origins and Properties of Multiple Imputation Methods for Missing Data

### 2.1.1 A Short History of Imputation Methods

Item missing data has long been recognized as a problem for data analysts. Early solutions to the problem of missing data were directed to specific distributions for the variables of interest and patterns of missing data. For example, Buck's (1960) method introduced imputations of conditional mean values for each pattern of missing observations in a multivariate normal vector of variables.

Broad, formal recognition of imputation as a statistical technique for dealing with missing data may have originated with the National Research Council's (NRC) Panel on Incomplete Data. Many of the earliest papers on imputation concepts and theory appear in the 1985 three-volume publication produced by the panel (Madow and Olkin 1983). Throughout the 1980s, statisticians continued to conduct research and to publish on imputation methods (Kalton 1983; Rubin 1980; Sande 1983). Of particular relevance to the software and methods described in this volume, the general theory and methods were greatly extended by the introduction of

the multiple imputation method (Rubin 1987). Despite these developments, the introduction of imputation methods to statistical practice at that time was a slow process and by no means universal.

Prior to the mid-1980s, the accepted procedure among most data analysts was to explicitly denote values as missing (e.g., the "." symbol for numeric data in SAS) but to take no corrective steps in actual data analysis other than to analyze complete data cases. During the 1980s, major federal survey programs in the United States and Canada took the lead in the development and application of basic imputation methods such as regression imputation and the hot deck imputation method. In the United States, developments in imputation methods were promoted by programs such as the Survey of Income and Program Participation (SIPP), programs that required the collection of many financial variables that were subject to significant rates of item missing data. By the early 1990s, large-scale, general purpose imputation of item-missing values in major survey data sets had become a common and accepted statistical practice.

During the 1990s and continuing to the present, the demand for practical methods to address increasingly large and complex missing data problems in surveys and other statistical investigations led to an explosion of new theoretical work during the next two decades, much of it focused on methods of multiple imputation (van Buuren 2012).

In the wake of these advances in imputation theory, general purpose procedures for multiple imputation of item missing data (PROC MI) and statistical analysis of imputed data sets (PROC MIANALYZE) were introduced to SAS and other major statistical software systems. Today, despite multiple imputation's roots in problems of missing data for large complex surveys, MI techniques and software are being applied to a wide range of statistical problems that involve missing data (Reiter and Raghunathan 2007).

As introduced in Chapter 1, multiple imputation in SAS is a three-step procedure for the treatment of missing data in statistical analysis. In SAS:

1. The analyst defines a multivariate "imputation model" for the data, and under this model PROC MI is used to independently impute missing values in the original data set $m$=1,…,$M$ times, generating $M$ complete "repetition" versions of the analysis data set.
2. Standard or SURVEY procedures (e.g., PROC MEANS/SURVEYMEANS, PROC REG/SURVEYREG, etc.) in SAS are then used to analyze each of the $M$ completed data sets and output the results.
3. PROC MIANALYZE inputs the results of the $M$ separate analyses and applies multiple imputation formulae to generate estimates, standard errors, confidence intervals, and test statistics for the descriptive statistics or model parameters of interest.

This chapter provides an intermediate-level introduction to multiple imputation theory and methods that we feel are most relevant to the SAS user. Readers who are interested in a more detailed theoretical review of MI are referred to Chapter 56 (PROC MI) and Chapter 57 (PROC MIANALYZE) of the SAS/STAT documentation or to one of the many excellent published texts on the subject (Allison 2001; Rubin 1987; Schafer 1997; van Buuren 2012).

### 2.1.2 Why the Multiple Imputation Method?

No imputation method or statistical modeling technique is optimal for all forms of missing data problems. As discussed in Chapter 1, there are at least six general approaches to the treatment of missing data in statistical analyses:

1. Conduct complete case analysis, ignoring cases with missing data (the list-wise deletion default in SAS);
2. Employ weighting of complete cases to compensate for missing data;
3. Employ full information maximum likelihood methods to analyze the data;
4. Analyze the incomplete data using the EM algorithm (Allison 2001; Little and Rubin 2002);
5. Perform single imputation of missing values using deterministic or stochastic approaches such as mean/mode imputation; regression imputation, predictive mean matching, nearest neighbor method, or

the hot deck (see Kalton and Kasprzyk [1986] and Little and Rubin [2002] for a comprehensive review);

6. Develop multiple imputations of the item missing data and employ MI estimation and inference in analysis.

Since the primary purpose of this volume is to instruct the data user in the capabilities of SAS for multiple imputation using the paired PROC MI and PROC MIANALYZE procedures, we will not go into depth on the particular and comparative properties of these missing data methods. Instead, we refer the reader to the many excellent references that are identified in the text or in the extended reference list. It is our view that the strengths of the multiple imputation approach to item missing data rest on the following attributes of properly designed and executed MI methods:

***MI is model-based.*** It ensures statistical transparency and integrity of the imputation process. To ensure robustness in analysis, the *imputation model* should be broader than the *analysis models* that will be analyzed using the imputed data (see Section 2.2). The model that underlies the imputation process is often an explicit distributional model (e.g., multivariate normal), but good results may also be obtained using techniques where the imputation model is implicit (e.g., nearest neighbor imputation).

***MI is stochastic.*** It imputes missing values based on draws of the model parameters and error terms from the predictive distribution of the missing data, $Y_{mis}$. For example, in linear regression imputation of the missing values of a continuous variable, the conditional predictive distribution may be: $\hat{Y}_{k,mis} = \hat{\beta}_0 + \hat{\beta}_{j\neq k} \cdot y_{j\neq k} + e_k$. In forming the imputed values of $y_{k,mis}$, the individual predictions incorporate multivariate draws of the $\hat{\beta}$ s and independent draws of $e_k$ from their respective estimated distributions. In a hot deck, predictive mean, or propensity score matching imputation, the donor value for $y_{k,mis}$ is drawn at random from observed values in the same hot deck cell or in a matched "neighborhood" of the missing data case.

***MI is multivariate.*** It preserves not only the observed distributional properties of each single variable but also the associations among the many variables that may be included in the imputation model. It is important to note that under the assumption that the data are missing at random (MAR), the multivariate relationships that are preserved are those relationships that are reflected in the observed data, $Y_{obs.}$

***MI employs multiple independent repetitions*** of the imputation procedure that permit the estimation of the uncertainty (the variance) in parameter estimates that is attributable to imputing missing values. This is variability that is in addition to the recognized variability in the underlying data and the variability due to sampling.

***MI is robust*** against minor departures from strict theoretical assumptions. No imputation model or procedure will ever exactly match the true distributional assumptions for the underlying random variables, *Y*, nor the assumed missing data mechanism. Empirical research has demonstrated that if the more demanding theoretical assumptions underlying MI must be relaxed that applications to data can produce estimates and inferences that remain valid and robust (Herzog and Rubin 1983).

***MI is very usable*** in real statistical applications, a practical feature that has been enhanced tremendously in the past ten years through the introduction of MI procedures to SAS and other major statistical software systems.

Setting aside for the moment its theoretical elegance and ties to sophisticated theory of Bayesian inference, the concept of multiple imputation was formulated by Rubin (1987) in large part to address the need for a robust method that could be applied to large data sets with many variable types. Furthermore, he recognized that in most cases the "data imputer" and the "data analyst" might well be different individuals with access to differing levels of information concerning the data collection design and the missing data process. Rubin sought a procedure that would enable the "imputer" to take full advantage of the available data and sophisticated imputation procedures yet leave the "data analyst" with a simple process to analyze the imputed data and obtain robust statistical inferences.

The most valid criticisms of the multiple imputation method (Fay 1996; Kim et al. 2006) have zeroed in on the notion that the imputer's statistical model for the imputation might be very different from the data models of interest to the many data analysts who will subsequently use the multiply imputed data. For example, in

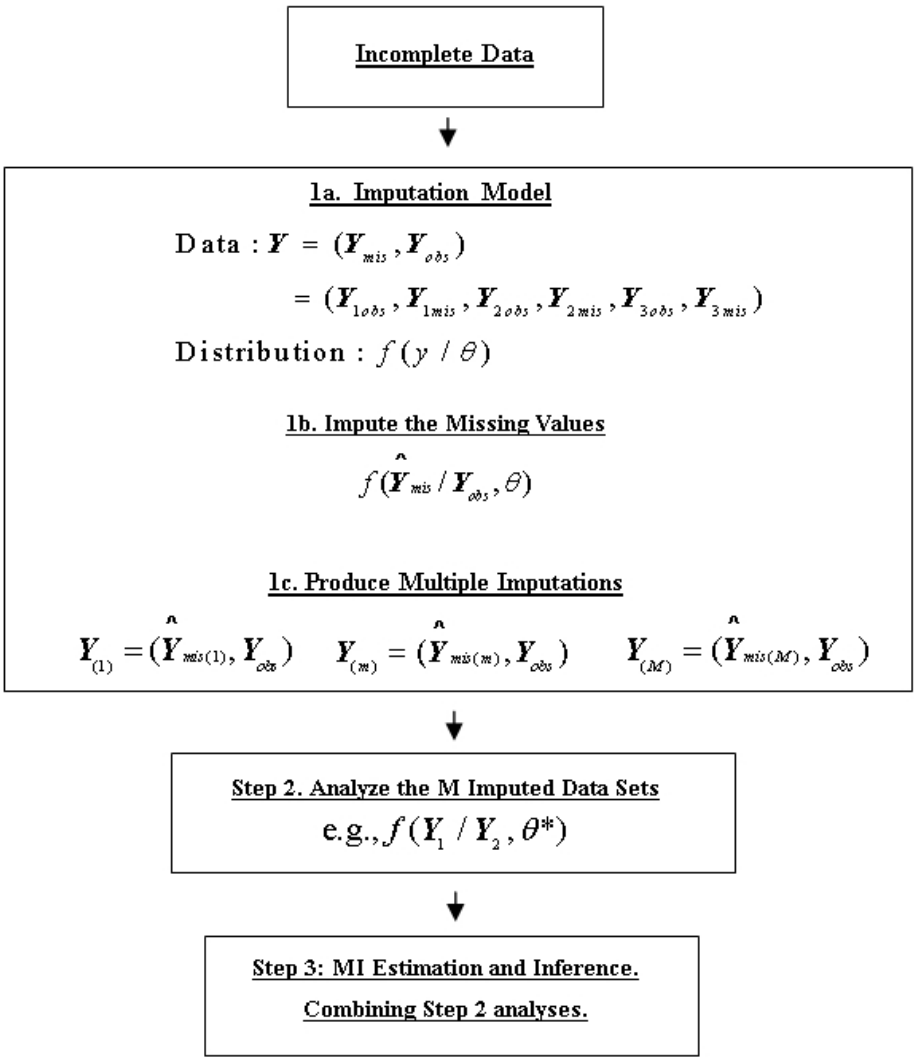
developing the multiple imputations of the item missing data, the imputer may fail to include an important analytical variable in her/his imputation model. Or, he/she may impute the missing data assuming a linear relation between two variables (e.g., height and weight) when, in fact, the relationship is nonlinear. The imputer may also fail to incorporate analytically important interactions or moderating effects among variables in the imputation model. Depending on the strength of these omitted effects and the amount of missing data for the variables in question, statistical analysis based on the imputed data may be subject to estimation and inferential bias. In Section 2.2 and throughout this volume, we will provide guidance on how to avoid such known pitfalls in the MI method.

In summary, for general missing data problems of the type covered in this book, PROC MI and PROC MIANALYZE provide a user-friendly platform for conducting a multiple imputation analysis that is theoretically-based, empirically-tested, and capable of generating robust statistical inferences.

### 2.1.3 Overview of Multiple Imputation Steps

Multiple imputation (MI) is not simply a technique for imputing missing data. It is also a method for obtaining estimates and correct inferences for statistics ranging from simple descriptive statistics to the parameters of complex multivariate models. As illustrated in Figure 2.1, a complete multiple imputation analysis in SAS can be organized into three sequential steps.

**Figure 2.1: Three Steps in the Multiple Imputation Approach (adapted with permission from Heeringa, West, and Berglund 2010)**




Step 1 defines the variables and distributional assumptions of the imputation model and applies specific MI algorithms to generate the multiple imputations of the missing values and output a completed data set for each of $m$=1,…,$M$ repetitions of the imputation process. The definition of the variables and distributional assumptions of the imputation model are primarily the responsibility of you, the user. The guidance and examples provided in this and later chapters are designed to help you make informed decisions on how to select the imputation model for your missing data problem.

SAS users have a wide array of choices for the imputation algorithm. As we will learn, the user-specified choice of the imputation algorithm will depend on the pattern of missing data (monotone, arbitrary), the type (continuous, binary, nominal, ordinal) of the variables to be imputed, and reasonable assumptions about the nature of the multivariate distribution of the variables specified in the imputation model. PROC MI will output the multiple copies of the completed data set as a "stacked" file with the _IMPUTATION_ variable added to distinguish the repetition assignment of each completed case.

Step 2 inputs the imputed data file produced by PROC MI and uses standard or SURVEY procedures in SAS with a BY _IMPUTATION_ statement to independently analyze each of the $m$=1,…,$M$ repetition data sets. A critical activity at this analysis step is to specify an appropriate output of the estimated statistics and their

standard errors from each of the repetitions of the analysis. The output data set of the estimated statistics and standard errors will be the required input for PROC MIANALYZE (step 3).

At Step 3, PROC MIANALYZE inputs the parameter estimates and standard errors from the preceding analysis step and applies the multiple imputation formulae to generate the MI estimate, standard errors, confidence intervals, and hypothesis test statistics for making inferences for descriptive statistics or model parameters.

The following three sections provide an intermediate-level introduction to these three steps of the MI process.

## 2.2 Step 1—Defining the Imputation Model

The actual process of imputing item missing values is governed by the imputation model, which we define as the set of variables that are available to the imputation process, $Y=\{Y_1,\ldots,Y_p\}$, and the distributional assumptions, $f(Y|\theta)$, for the multivariate relationships among the elements of $Y$.

Consider a trio of variables, $Y=\{Y_1$=diastolic blood pressure (mm Hg); $Y_2$=age (years); and $Y_3$= Body Mass Index (kg/m$^2$)$\}$ from the National Health and Nutrition Examination Survey (NHANES) 2009–2010 medical examination component (MEC). The data are restricted to NHANES respondents age 20 and older (n=6,059). The unweighted missing data rate is 8.4% for diastolic blood pressure (DBP) and 1% for body mass index (BMI), and following NHANES editing, age is observed for every case. The pattern of missing data is not monotone. One possible imputation model would be to include all three variables in the imputation process and to assume that the joint distribution for these three continuous variables is multivariate normal with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, that is, $f(\boldsymbol{y}|\boldsymbol{\theta}) = \text{MVN}(\boldsymbol{\mu}, \Sigma)$. Under this distributional model, multiple imputations of the missing data for diastolic BP and BMI are easily performed using the PROC MI Markov chain Monte Carlo (MCMC) method (Schafer 1997). The imputations performed under this imputation model would serve for univariate MI estimation of the mean diastolic BP, $\mu_1$, or mean BMI, $\mu_3$, and they would also serve for the MI estimation of multiple regression parameters, for example, the regression of diastolic BP on age and BMI.

### 2.2.1 Choosing the Variables to Include in the Imputation Model

The choice of variables to include in the imputation model should not be limited to only variables that have item missing data or variables that are expected to be used in a subsequent analysis. As a general rule of thumb, the set of variables included in the imputation model for an MI analysis should be much larger and broader in scope than the set of variables required for the analytic model. For example, if age and BMI are the chosen predictors in the analytic model for diastolic blood pressure, the imputation of item missing data for diastolic blood pressure and BMI might include many additional variables, such as gender, race/ethnicity, marital status, height, weight, systolic blood pressure, and so on. If the relationship of age to diastolic blood pressure is not linear, the regression model that is used to impute missing blood pressure measurements may include both a linear and quadratic term for age. If gender moderates the effects of age on diastolic blood pressure, an interaction term for age and gender should be included in the imputation model. Obviously, it is not feasible to define an imputation model and perform multiple imputations using every possible variable in the survey data set. Based on recommendations from Schafer (1999) and van Buuren (2012), some practical guidelines for choosing which variables to include in the imputation model are the following:

1. Include all key analysis variables: (dependent: $Y_1$ and independent: $Y_2$)
2. Include other variables that are correlated or associated with the analysis variables: ($Y_3$ )
3. Include variables that predict item missing data on the analytic variables: ($Z$)

Failure to include one or more analysis variables ($Y_1$ and $Y_2$) in the imputation model can result in bias in the subsequent MI estimation and inference. Including additional variables, ($Y_3$ ), that are good predictors of the analytic variables improves the precision and accuracy of the imputation of item missing data. Under the assumption that item missing data is MAR, incorporating variables ($Z$) that are correlated with the variables that have missing data and predict the propensity for response will reduce bias associated with the item missing data mechanism.

For multiple imputation, a general piece of advice for practitioners that has come from extensive empirical work and simulation testing is: "When in doubt, including more variables in the imputation model is better."

### 2.2.2 Distributional Assumptions for the Imputation Model

Statistical models that define the relationships of the variables that are jointly considered in the missing data problem are key to all imputation methods. Under some imputation methods such as hot deck imputation the models are implicit in the mechanics of the procedure. Other imputation methods are based on explicit probability models, $f(Y \mid \theta)$, for the multivariate relationships among the elements of the complete data *Y*. In theoretical discussions of multiple imputation methods, convenient choices of a multivariate model for the joint distribution of the broad set of imputation variables might be multivariate normal (continuous) or multinomial (classification). As described in Section 2.3, several of the imputation options in PROC MI explicitly assume that the imputation variables follow one of these standard multivariate data models. Other procedures such as the fully conditional specification (FCS) method will not specify the distributional models for the data but use iterative simulation to approximate the complex and unknown distribution for problems involving a mixture of variable types and arbitrary patterns of missing data.

## 2.3 Algorithms for the Multiple Imputation of Missing Values

Once the user has defined the imputation model, an imputation algorithm is used to generate $m$=1,…,$M$ completed data sets in which the missing values**,** $Y_{mis}$, have been imputed. In this volume, the $m$=1,…,$M$ independently imputed versions of the data set are termed repetitions.

### 2.3.1 General Theory for Multiple Imputation Algorithms

The theoretical development of multiple imputation methods for missing data is rooted in the Bayesian framework for statistical inference. Within this framework, the task of imputing missing values, $Y_{mis}$, in a data set equates to a random draw of an imputed value from the posterior predictive distribution of the missing data which we will denote as $f(Y_{mis}|Y_{obs},\theta)$ where $\theta$ is the vector of parameters (e.g., $\mu$,$\Sigma$) or functions of these parameters (e.g., regression parameters $\beta$, $\sigma_{y.x}$) that uniquely define this predictive distribution for the missing values. Fortunately for most of us, we do not need to be experts in Bayesian inference to apply multiple imputation methods in practice. However, it is useful to have a simple overview to understand how this "posterior predictive distribution" is actually derived or simulated under the various MI techniques that are included in SAS PROC MI.

To avoid confusion in terms, from this point forward we will refer to the posterior predictive distribution as the predictive distribution for the missing data values. Examining the notational symbol for this distribution, $f(Y_{mis}|Y_{obs}, \theta)$, we see that this distribution is a function of our observed data, $Y_{obs}$ and distributional parameters, $\theta$. The process of imputing missing values for $Y_{mis}$ requires that we first derive the predictive distribution, $f(Y_{mis}|Y_{obs}, \theta)$—labeled the "posterior" or "P" step—and then make random draws of imputed values from the predictive distribution, $Y_i^* \approx f(Y_{mis} \mid Y_{obs}, \theta)$. The drawing of random variates from the predictive distribution for $Y_{mis}$ to fill in the missing values in the data is called the imputation or "I" step in the imputation process. In many imputation procedures, the process of generating "draws" from the predictive distribution is based on familiar predictions from regression functions (linear, logistic, discriminant function) that impute the missing value based on its expected relationship to other observed covariates.

We noted above that the predictive distribution $f(Y_{mis}|Y_{obs}, \theta)$ is a function of our observed data, $Y_{obs}$, and distributional parameters, $\theta$. The data, $Y_{obs}$, however, as in any problem of statistical inference from a sample of data, has values of the parameters, $\theta$, that must be estimated or derived. The same is true for the P-step in the imputation process. In the Bayesian framework for estimation and inference, these distributional parameters, e.g., $\theta=\{\mu, \Sigma\}$ for the multivariate normal, are assumed to have a prior probability distribution, $g(\theta)$. The observed data are used to update our information about the likely values of the true $\theta$, to produce a new posterior distribution for the parameters, $p(\theta|Y_{obs})$. To execute the P-step in the imputation process, it is necessary to derive this posterior distribution for $\theta$ exactly or to somehow simulate it closely through an iterative computational process. If the form of the complete data distribution, $f(Y|\theta)$, and the prior distribution,

*g(θ)*, are specified, the form of the posterior distribution of the parameters can often be derived through a formula based on Bayes' Rule:

$$p(\theta \mid Y_{obs}) = \frac{f(y \mid \theta) \cdot g(\theta)}{\int_{\theta} f(y \mid \theta) \cdot g(\theta) \cdot d(\theta)}$$

The monotone methods in PROC MI (linear regression for continuous *Y*, logistic regression for binary and ordinal *Y*, discriminant function method for nominal categorical *Y*) impute one variable at a time. The algorithm for each of these three MONOTONE methods imputes missing data based on a known expression for the posterior distribution of the parameters in the predictive distribution of the missing data. See the SAS/STAT PROC MI documentation for details.

In the common situation where the missing data problem is multivariate, has an arbitrary pattern of missing values, and may include variables of differing type (continuous, nominal, binary, ordinal), it is analytically difficult or impossible to evaluate the true expression for the joint posterior distribution, $p(\theta|Y_{obs})$. In such cases, statisticians have devised iterative simulation techniques that permit us to approximate draws from the analytically intractable complex joint posterior. The PROC MI MCMC method is such an algorithm to simulate the joint posterior, $p(\theta|Y_{obs})$, for arbitrary data patterns in which the underlying complete data are assumed to follow a multivariate normal distribution. The FCS method uses an iterative sequence of draws from conditional distributions (linear regression or predictive mean matching for continuous *Y*, logistic regression for binary and ordinal *Y*, discriminant function method for nominal categorical *Y*) to simulate draws from the highly complex joint posterior distribution of parameters for a set of variables of mixed distributional type.

To summarize, depending on the pattern of missing data and variable types, PROC MI provides three primary classes of methods for generating the multiple imputations. If the pattern of missing data is univariate or monotonic (Figure 2.2), the monotone option is the method of choice. For an arbitrary multivariate pattern of missing data (Figure 2.3), the choice is between the MCMC or the FCS methods. Table 2.1 summarizes methods available in SAS (v9.4) according to the pattern of missing data and the type of variable being imputed.

**Table 2.1: SAS PROC MI Imputation Methods**

| Missing Data Pattern | Variable Type | Method |
|---|---|---|
| Monotone | Continuous | Linear regression, predictive mean matching, propensity score |
| | Binary/ordinal | Logistic regression |
| | Nominal | Discriminant function |
| Arbitrary | Continuous | With continuous covariates:<br>MCMC monotone method<br>MCMC full-data imputation |
| | Continuous | With mixed covariates:<br>FCS regression<br>FCS predictive mean matching |
| | Binary/ordinal | FCS logistic regression |
| | Nominal | FCS discriminant function |

The following paragraphs provide an overview of each of these algorithms as they apply to variables of differing types.

### 2.3.2 Methods for Monotone Missing Data Patterns

The PROC MI algorithm for the multiple imputation of monotone missing data involves a non-iterative sequence of steps to generate each of the $m$=1,…,$M$ repetition imputations of $Y_{mis}$. To outline the steps in the algorithm, we will use the notation and missing data pattern shown in Figure 2.2 with $Y$={$Y_1$, $Y_2$, $Y_3$, $Y_4$, $Y_5$}. In Figure 2.2, $Y_1$ and $Y_2$ are shown as fully observed—there are no missing values for these two variables. Moving from left to right in the ordered data array, the remaining variables have increasing amounts of missing data, and the missing data is always nested so that whenever $Y_3$ is missing, $Y_4$ and $Y_5$ are missing as well. Likewise, any cases that are missing values on $Y_4$ are also missing $Y_5$.

**Figure 2.2 Monotone Multivariate Missing Data Pattern**

| | Variables | | | | |
|---|---|---|---|---|---|
| Obs | Y1 | Y2 | Y3 | Y4 | Y5 |
| 1 | | | ? | ? | ? |
| 2 | | | | ? | ? |
| 3 | | | | | ? |
| 4 | | | | | ? |
| 5 | | | | | ? |

In this example, the sequence of imputations in the monotone pattern therefore begins with imputation of missing values of $Y_3$.

The P-step in the imputation of missing $Y_3$ will utilize the relationship of the observed values of $Y_3$ to the corresponding observed values of $Y_1$ and $Y_2$ to estimate the parameters of the predictive distribution, p($Y_{3,mis}$|$Y_1$,$Y_2$, , $\theta_3$). The predictive distribution and the parameters to be estimated will depend on the variable type for $Y_3$. PROC MI will use either linear regression or predictive mean matching (continuous), logistic regression (binary or ordinal categorical), or the discriminant function method (nominal categorical) to estimate the predictive distribution. For example, if $Y_3$ is a continuous scale variable, the default predictive distribution is the linear regression of $Y_{3,obs}$ on $Y_1$, $Y_2$ with parameters, $\theta_3$ ={β-the vector of linear regression coefficients, and $\sigma^2_3$ the residual variance}. To ensure that all sources of variability are reflected in the imputation of $Y_{3,mis}$, the values of the parameters for the predictive distribution, p($Y_{3,mis}$|$Y_1$,$Y_2$, $\theta_3$), are randomly drawn from their estimated posterior distribution, p($\theta_3$| $Y_1$,$Y_2$).

#### Linear Regression

In PROC MI, when the monotone (or FCS) imputation method is specified, the default imputation method for continuous variables is linear regression. Here we will illustrate the P-step and I-step for the monotone missing data pattern. In FCS, the P-step and I-step are similar with minor adaptations to account for the iterative cycles of the algorithm.

Assume that $Y_3$ in our example is a continuous variate with missing values. The MONOTONE method will first regress the observed values of $Y_3$ on the values of the more fully observed {$Y_1$,$Y_2$} using the standard model:

$$Y_3 = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \varepsilon$$

The regression will yield current estimates of the regression parameters, $\hat{\beta} = \left\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right\}$, and, the residual variance $\hat{\sigma}_3^2$, and also $V_{3,}$, the inverse of the sum of squares and cross products (SSCP) matrix from the regression of $Y_3$ on $Y_1$ and $Y_2$. These regression estimates define the posterior distribution for the regression model parameters.

Based on the estimated posterior for the regression model parameters, the **I-step** imputes "draws," $Y_3$*. The first step in this imputation is to draw random parameter values from their joint posterior distribution. First, the value for the residual variance is drawn from its posterior (assuming a non-informative prior):

$$\sigma^2_{*j} = \hat{\sigma}^2_j (n_j - k - 1) / g$$

*where* :

$n_j$ = count of non-missing values for variable being imputed;

$k$ = the number of parameters (excluding intercept) in the model;

g= a random draw from the central Chi-square distribution, $\chi^2_{n_j - k - 1}$ .

In our example of imputing missing values of $Y_3$, $j$ is 3 and $k$=2. Conditional on the drawn value of $\sigma^2_{*j}$, draws of the regression parameters are made from their conditional posterior distribution:

$$\beta_* = \hat{\beta} + \sigma_{*j} \sqrt{V}' Z$$

*where* :

$\sqrt{V}$ = the upper triangle in the Cholesky decomposition (square root) of $V = (X'X)^{-1}$; and

Z = k + 1 dimensional vector of independent normal, N(0,1), variates.

Finally, the actual imputations of missing values of $Y_3$ are derived from the regression equation:

$$Y_{3*} = \beta_{0*} + \beta_{1*} Y_1 + \beta_{2*} Y_2 + z\sigma_{*3}$$

*where* :

$z$ = a random draw of a standard normal, N(0,1), deviate.

The algorithm will then turn to the P-step for $Y_4$, using linear or logistic regression or discriminant classification to define $p(Y_{4,mis}|Y_1, Y_2, Y_{3,obs}\ Y_{4,obs}\ , \theta_4)$. Draws from the corresponding predictive distribution for $Y_4$ will be used to create the imputations, $Y_{4*}$.

The imputation of this monotone sequence will conclude by applying the process described for $Y_{4,mis}$ to the missing values for the fifth and final variable, $Y_{5,mis}$.

Each step in the sequence is a univariate imputation of a single variable that is conditioned on the more completely observed variables in the monotone pattern.

## Predictive Mean Matching

Predictive mean matching (PMM) is an option in PROC MI for the imputation of missing values for continuous variables. The PROC MI predictive mean matching method for imputing continuous variables utilizes the same initial steps as the linear regression method above; however, the I-step differs. Using our example of imputing a missing value for a continuous $Y_3$, the draws of the regression parameters are used to construct a regression prediction of $Y_3$ for each case.

The data set is sorted according to regression predictions of Y3 for both the missing and observed cases. A "neighborhood" for each missing value is defined by the set of observed cases that most closely matches on the regression prediction. Each missing value of Y3 is then imputed by drawing and substituting an observed value that is in the "neighborhood" of its predicted value. The size of the neighborhood used to select the random donor is controlled by the imputer. The default in PROC MI is to select the imputation donor from the nearest K=5 cases. Note that the random draw of an actual observation ensures that the imputation will lie within the range of actual observed values. This algorithm for predictive mean matching imputation for a continuous variable is also an option in the FCS method.

## Logistic Regression

The logistic regression method is used in the MI monotone and FCS methods to impute binary or ordinal classification variables. The actual process of imputing missing values under this model follows a P-step, I-step sequence that is similar to that detailed above for the linear regression methods. In the P-step, logistic regression is applied to observed values of the dependent variables (e.g., $Y_4$ in our five-variable example) and observed predictors (e.g., $Y_1$, $Y_2$, $Y_3$ ), yielding the fitted logistic model for the probability that the missing value belongs to each category of the binary or ordinal classification variable. To illustrate, assume that $Y_4$ is a binary variable with values (0,1). The fitted logit model with the DESCENDING option models the logit of the probability that $Y_4$ =1. The fitted model is:

$$\text{logit}(p(Y_4 = 1)) = \log\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 Y_1 + \hat{\beta}_2 Y_2 + \hat{\beta}_3 Y_3$$

The fitted logistic regression model yields estimates of the logistic regression parameters, $\boldsymbol{\beta}=\{\beta_0, \beta_1, \beta_2, \beta_3\}$ and the covariance matrix for parameters **V**. As in linear regression, the posterior distribution of the logistic regression parameters is assumed to be multivariate normal and random draws from this posterior based on the following:

$$\beta_* = \hat{\beta} + \sqrt{V}' Z$$

*where* :

$\sqrt{V}$ = the upper triangle in the Cholesky decomposition (square root) of the Covariance Matrix for the $\hat{\beta}$; and

$Z$ = k + 1 dimensional vector of independent normal, N(0,1), variates.

To complete the P-step for imputing missing values of the binary variable, the inverse logit transform is computed using the drawn values of the $\beta_*$ s:

$$p(Y_4 = 1) = \frac{\exp(Y_{j<4}\beta_*)}{1 + \exp(Y_{j<4}\beta_*)}$$

In the I-step, a random number drawn from a uniform (0,1) probability distribution is used to determine the imputed category (e.g., $Y_4$=1 or $Y_4$=0 for binary $Y_4$). To understand how this works, consider the example of a binary measure of self-reported type-2 diabetes status. Assume that the P-step logistic model predicts the probability of a "Yes" response for a missing value to be: p(Yes)=0.12. If the drawn $U(0,1)$ random number for the imputation of the missing observation is $u$=0.73, since the random drawn is greater than the predicted probability ($u$ >p(Yes)), a value of "No" will be imputed to the case. If, instead, the random number draw was $u$=0.08, the random number is less than or equal to the predicted probability and a value of "Yes" would be imputed. In applications of the MONOTONE and FCS methods (discussed below), the logistic regression imputation method is the technique used to impute missing values of both binary and ordinal variables.

The process illustrated here for a binary variable is easily extended to the ordinal variables with $K$>2 levels.

## Discriminant Function Method

In major software systems, multiple imputation of missing data for a classification variable, $Y_j$, with nominal category groups, $g$=1,…,$G$ (e.g., workforce status), is typically performed using one of two methods: a discriminant function method or a method based on the generalized (multinomial) logit model. Under both the monotone and FCS methods, PROC MI uses the discriminant function method to impute missing values for nominal classification variables. The discriminant function method and the generalized logit approach each employ a P-step that simulates the parameters (group probabilities) of the multinomial posterior distribution for $Y$. Both methods relate these multinomial class probabilities to a vector of observed covariates, X={$X_1$, **…,**$X_p$}. The discriminant function method requires a more rigid set of distributional assumptions concerning these covariates. Within each category group, $g$=1,…,$G$, the vector of observed covariates, $X_g$, is assumed to be

approximately multivariate normal, and furthermore the variance-covariance matrix of these normally distributed covariates is assumed to be constant across groups, that is, $\Sigma_g=\Sigma$ for all groups.

Given a multivariate normal vector of covariates for a case, the predictive distribution of the missing value for the case is the multinomial distribution with posterior probability of belonging to category, $g$=1,…,$G$ obtained from the following formula:

$$p(Y_{i,mis} = g \mid x_i) = \frac{\exp(-0.5 \times D_g^2(x_i))}{\sum_c \exp(-0.5 \times D_g^2(x_i))}$$

with :

$D_g^2(x_i)$ = the squared distance discriminant function value for case i and group g ,

$= (x_i - \mu_{*g})' \Sigma_*^{-1} (x_i - \mu_{*g}) - 2\log(q_{*g})$

where :

$x_i = \{x_{i1}, ..., x_{ip}\}$ vector of covariates for case i;

$\mu_{*g}$ = draw of posterior mean vector for X in group g;

$\Sigma_*^{-1}$ = draw of posterior covariance matrix for X (common to all groups); and

$q_{*g}$ = draw of prior probability of group $g$ = 1, ..., $G$ membership.

Before these multinomial probabilities can be evaluated in PROC MI, values of the parameters $\Sigma_*$, $\mu_{*g}$, and $q_{*g}$ must be drawn from the appropriate distributions. The sequence of three "draws" begins with $\Sigma_*$, the simulated posterior value of the common MVN variance covariance matrix, Σ. Through the PCOV= option, PROC MI provides two options. The default (also PCOV=POSTERIOR) is to draw the value of $\Sigma_*$ from its posterior distribution, which under the PROC MI assumption of a noninformative prior is the inverted Wishart distribution:

$p(\Sigma \mid X) \sim W^{-1}(n - G, (n - G)S)$,

where :

n = total sample size;

G = total number of groups in nominal classification variable Y; and

S = the fixed estimate of the pooled covariance matrix for the observed Xs,

$$= \frac{1}{(n - 1)} \sum_g (n_g - 1) S_g$$

The option, PCOV=FIXED, specifies that the $\Sigma_*$ is simply set to the fixed value of S estimated from the observed X values.

The second step is to draw the values of the group-specific mean vectors from their respective posterior distributions. PROC MI assumes a noninformative prior for the values of these group mean vectors. The posterior distribution for each group mean is therefore the normal distribution:

$$p(\mu_g \mid \Sigma, \bar{x}_g) \sim N[\bar{x}_g, (1 / n_g)\Sigma_*]$$

Note that these draws condition on the previously drawn value for $\Sigma_*$.

Next, PROC MI computes or draws the q values of the prior probabilities of belonging to group $g$=1,…,$G$. The PRIOR= option allows the user to control the specification of the prior distribution for these probabilities. The default option is PRIOR=JEFFRIES, a noninformative Dirichlet prior (not shown).

With the $\Sigma$, $\mu$, and $q$ draws completed, the discriminant function imputation method uses the formula given above to estimate the posterior probabilities of group membership of each missing data case. Across the *G* categories of the variable, the sum of these posterior probabilities will be 1.0 for each case. The remaining step in the imputation process is to use these probabilities to assign the missing *Y* value to one of the *G* nominal groups. As is the case for the logistic regression method, the I-step in the discriminant function method uses a random draw from a *U*(0,1) distribution. This *U*(0,1) random number draw is then compared to the accumulation of the *g*=1,…,*G* category probabilities from the discriminant classification model to determine the category for the imputed value, *Y**.

### Propensity Score

PROC MI also offers a propensity score option (Schafer 1999) for performing imputation of missing data. This is a univariate method that was developed for use in very specialized missing data applications. It does not incorporate or preserve associations among the variables in the imputation model, and therefore we feel should not be recommended for MI applications where multivariate analysis is the ultimate goal. For this reason, we encourage analysts to use one of the other options available in PROC MI.

## *2*.3.3 Methods for Arbitrary Missing Data Patterns

Here, we consider the more typical data context where the imputation model is multivariate, includes variables of all types, and has an arbitrary pattern of missing data (Figure 2.3). In such cases of a "messy" pattern of missing data where exact methods do not strictly apply, the authors of multiple imputation software have generally followed one of three general approaches. Each of these three approaches to an arbitrary pattern of missing data are available in PROC MI.

**Figure 2.3: Arbitrary Multivariate Missing Data Pattern**

| | Variables | | | | |
|---|---|---|---|---|---|
| Obs | Y1 | Y2 | Y3 | Y4 | Y5 |
| 1 | | | ? | ? | |
| 2 | ? | ? | | ? | ? |
| 3 | ? | | ? | | |
| 4 | | ? | | | |
| 5 | ? | | | ? | |

### The Markov chain Monte Carlo (MCMC) Method: Using an Explicit Multivariate Normal Model and Applying Bayesian Posterior Simulation Methods

A first algorithmic approach to generating imputations for an arbitrary pattern of item missing data is to declare an explicit probability model for the data and a prior distribution for the parameters, $g(\theta)$ (Schafer 1997). The earliest versions of multivariate multiple imputation programs for continuous data (Schafer 1999) assumed a multivariate normal distribution for all variables, $f(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \mathrm{MVN}(\boldsymbol{\mu}, \Sigma)$, and a noninformative or Jeffries prior distribution for the parameters $\mu$ and $\Sigma$. In the case of complete data, the posterior distribution $p(\mu,\Sigma|Y)$ can be derived under Bayes' Rule. However, for an arbitrary pattern of missing data where individual cases are missing different combinations of the variables in $Y=\{Y_1,\ldots,Y_p\}$, the same posterior—now conditional only on the observed data—is difficult or impossible to derive in a closed form. The PROC MI MCMC full-data imputation method uses an iterative Markov chain Monte Carlo method to simulate draws from the posterior, $p(\mu,\Sigma|Y_{obs})$. For the curious reader, Schafer (1999) provides a detailed description of the MCMC algorithm. Here, we will describe the algorithm in general terms. The MCMC algorithm involves an iterative sequence of paired I-steps and P-steps.

#### I-Step

At each iteration of the simulation ($t$=1,…,$T$), the MCMC algorithm draws imputations from the current iteration's predictive distribution, $f(Y^{(t+1)}_{mis}|\ Y_{obs}, \mu^{(t)}, \Sigma^{(t)})$. The imputation proceeds case by case, taking into account the pattern of missing variables for the case. For example, the predictive posterior for a case with the

observed/missing pattern of $Y_i=(Y_1, . , Y_3, ., Y_5)$ is different from that for $Y_i=(Y_1, . , . , Y_4, Y_5)$. For efficiency, MCMC uses the SWEEP operator (Goodnight 1979)—a computationally convenient way to estimate linear regression parameters from $\Sigma^{(t)}$—to derive the conditional distributions needed to simulate the predictive posterior for each possible pattern of missing data.

### P-Step

After each I-step, the parameter values for the predictive distribution are updated by draws from the completed data posterior, $p(\mu,\Sigma|Y_{obs},Y^{(t+1)}_{mis})$.

In theory, if the chain of MCMC I-step/P-step pairs is allowed to continue for many iterations, the algorithm will converge so that the imputation draws for the missing values will simulate draws from the true joint posterior, $p(Y_{mis}|\mu,\Sigma, Y_{obs})$. Once a sufficient burn in period of iterations has passed, the $m$=1,…,$M$ repetitions can be taken as a successive series of systematic draws in the single imputation chain. Another option is to use $m$=1,…,$M$ MCMC runs in parallel chains and obtain each repetition of the multiple imputation as single draws from each of the independent MCMC chains. The single chain option is the default in SAS, but users may optionally request a multiple chain approach in applications of the MCMC method.

There is no exact test to establish that the MCMC algorithm has in fact converged to the joint posterior distribution. Whether single or multiple MCMC chains are used, the length of the burn in period (or initial iteration cycles) that is required to ensure convergence of the MCMC algorithm will depend on the number of variables and the rates and patterns of missing data. PROC MI defaults to NBITER=200, meaning 200 burn-in iterations. In addition, SAS provides several graphic tools to evaluate the convergence of the MCMC algorithm. Trace plots permit the user to plot the value of posterior estimates of means and variances against the iteration number. These plots should show the posterior mean and variance for the $M$ multiple imputation repetitions converging to a stationary distribution as the number of iterations increases. Ultimately, if convergence is reached, the plotted posterior mean and variance should vary randomly. After a sufficient number of burn-in iterations, the trace plots for the posterior mean and variance should not exhibit any patterns or trends either within or across the traces for the independent MI repetitions. Autocorrelation plots graph the autocorrelation in the values of the posterior parameters as a function of the "lag" in the number of iterations. Autocorrelation plots should show the lagged autocorrelations decline in value, ultimately varying randomly about zero.

The MCMC algorithm makes the assumption that the underlying variables in the imputation model are distributed as a multivariate normal random variable, $Y$~MVN($\mu$,$\Sigma$). In the case of continuous variables that are highly skewed or otherwise non-normal in distribution, PROC MI currently enables the user to specify transformations (e.g., a natural log transformation for a log normally distributed income measure).

PROC MI also allows the user to force the MCMC assumptions on data of mixed type and then use post-imputation rounding to restore the imputed variables to their original measurement scale. For example, a binary variable imputed as 1.15 would be rounded to "1," while a value of 1.73 would be rounded to "2." We agree with Allison (2005) and do not recommend the use of MCMC with the rounding technique for imputing classification variables. With the availability of the FCS method for imputation of mixed variable types, we advise the user to use imputation methods (regression, logistic regression, discriminant function method) that are directly appropriate to the variable type.

## Transform the Arbitrary Missing Data Pattern to a Monotonic Missing Data Structure

A second approach to dealing with arbitrary missing data is to transform the pattern of item missing data to a monotonic pattern by first using simple imputation methods or an MCMC posterior simulation approach to fill in the missing values for variables in the model that have very low rates of item missing data. If all variables in the imputation model are assumed to be continuous, the MCMC method with the MONOTONE option can be used to implement this approach. As previously described, imputation of a true monotonic pattern of item missing data is greatly simplified—reduced to a sequence of imputations for single variables. This MCMC monotone approach works best when the generalized pattern of missing data is dominated by missing data for one or two variables.

Consider the missing data for the variables selected for analysis of a Health and Retirement Study (HRS) survey question on serious falls in the past two years shown in Table 2.2. The generalized pattern of missing data is

dominated by the missing data on falls (4.5%), with a lesser rate for weight (1.4%), virtually no missing data for arthritis (0.2%), and complete data for age and gender. Under the MCMC MONOTONE option, PROC MI uses MCMC to impute the minimum number of missing data values to transform the problem to a monotone missing pattern. Noniterative monotone regression or predicted mean matching imputations can then be used to sequentially fill in the remaining missing values. This technique will be illustrated through a worked example in Chapter 6.

**Table 2.2: Item Missing Data Rates for Variables Included in the 2006 HRS Falls Model (n = 11,731 Eligible Respondents Age 65 and Older)**

| Variable | Falls | Age | Gender | Arthritis | Weight |
|---|---|---|---|---|---|
| % Missing | 4.50% | 0.00% | 0.00% | 0.20% | 1.40% |

### FCS, Sequential Regression, and Chained Regressions

The FCS approach is the third alternative to multiply impute arbitrary missing data for large mixed sets of continuous, nominal, ordinal, count, and semicontinuous variables. The FCS method is also labeled the sequential regression algorithm (Raghunathan et al., 2001) or the "chained equations" approach (Carlin, Galati, and Royston 2008; Royston 2005; van Buuren, Boshuizen, and Knook 1999). Each of these algorithms is based on an iterative algorithm. Each iteration ($t$=1,…,$T$) of the algorithm moves one by one through the sequence of variables in the imputation model, for example, $Y=\{Y_1,Y_2,Y_3,Y_4,Y_5\}$ as illustrated in Figure 2.3. At each iteration and for each variable, there is a P-step and an I-step. In the P-step, the current (iteration $t$) values of the observed and imputed values for the imputation model variables are used to derive the predictive distribution of the missing values for the target variable. To model the conditional predictive distribution of individual $Y_k$, $f(\hat{Y}_k^{(t)} / \hat{\boldsymbol{Y}}_{j\neq k}^{(t)}, \boldsymbol{\theta}^{(t)})$, PROC MI uses the same regression or discriminant function methods described above for the monotone missing data patterns. That is, linear regression or the regression-based predicted mean matching (PMM) approach is used to impute missing values for continuous variables, ordinal logistic regression to generate imputations for binary or ordinal classification variables, and the discriminant function method for nominal classification variables. Updated imputations, $\hat{\boldsymbol{Y}}_k^{(t)}$, are then generated by stochastic draws from the predictive distribution defined by the updated regression model. When the last variable in the sequence has been imputed, the algorithm cycles again through each variable, repeating the chain of regression estimation and imputation draw steps. The burn-in iteration of the cycles continues until the user-defined algorithm convergence or system default value is met (i.e., NBITER=x iterations or the default of NBITER=10, as specified in PROC MI).

Under an explicitly defined imputation model, $f(Y|\theta)$, and a suitable prior distribution, $g(\theta)$, for the distributional parameters, this FCS algorithm will, in theory, converge to the joint posterior distribution. Since the sequential regression method never explicitly defines $f(Y|\theta)$, the assumption must be made that the posterior distribution does exist and that the final imputations generated by the iterative algorithm do represent draws from an actual, albeit unknown, joint posterior distribution. Although the exact theoretical properties of the resulting FCS imputations remain somewhat of an unknown, in most applications the sequential regression approach does converge to a stable joint distribution. The multivariate imputations generated by the algorithm show reasonable distributional properties and have empirically been shown to produce results comparable to those for the EM algorithm and exact methods of Bayesian posterior simulation (Heeringa, Little, and Raghunathan 2002).

## 2.4 Step 2—Analysis of the MI Completed Data Sets

Although there is no requirement to separate the MI analysis step from the estimation and inference step, for most analyses SAS has chosen to split these two steps. This provides the user maximum flexibility to use the wide array of existing SAS standard and SURVEY procedures to conduct analysis. The input to the analysis step is the concatenated or "stacked" data set produced by PROC MI. The analyst then specifies standard or SURVEY procedure statements with a BY _IMPUTATION_ statement to independently analyze each of the $m$=1,…,$M$ repetition data sets.

A critical activity at this analysis step that will be carefully covered in examples in following chapters is to specify an appropriately structured output data set containing the estimated statistics and their standard errors from each of the *M* repetitions of the analysis. The output data set of the estimated statistics and standard errors will be the required input for PROC MIANALYZE (step 3). Chapters 4 through 8 provide examples of how to ensure that output of parameters estimates and variance/covariance matrices from SAS analysis procedures is correctly formatted for input to PROC MIANALYZE.

## 2.5 Step 3—Estimation and Inference for Multiply Imputed Data Sets

Once SAS procedures have been used to individually analyze the multiply imputed data sets and save the results in an output file, the next step in a complete MI analysis is to compute multiple imputation estimates of the descriptive statistics or model parameters and the variance of these MI estimates. The MI estimates and standard errors can then be used to construct confidence intervals for population quantities or model parameters.

Many of the major software systems include a pair of programs to conduct multiple imputation analysis—one program to perform the multiple imputation of item missing data and a second to perform MI estimation and inference. In SAS, as previously discussed, this pair of programs is PROC MI and PROC MIANALYZE. Although SAS supports coordinated processing of both the imputation and estimation/inference phases of an MI analysis, provided some care is taken in choosing the imputation model, the imputation and estimation phases can be performed separately by different persons (i.e., imputation by data producers, analysis by data users) or using separate programs (e.g., imputations in IVEware [Raghunathan, Solenberger, and Van Hoewyk 2002], analysis of completed data repetitions using SAS procedures and MI estimation and inference in PROC MIANALYZE).

### 2.5.1 Multiple Imputation—Estimators and Variances for Descriptive Statistics and Model Parameters

Following Rubin (1987), multiple imputation estimates of descriptive statistics and model parameters are computed by simply averaging the estimates from the $m$=1,…,$M$ independent repetitions of the imputation algorithm:

$$\overline{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$$

where $\hat{\theta}_m$ = estimate of $\theta$ from the completed data set $m$=1,...,$M$.

Rubin (1987) also proves that the corresponding multiple imputation variance for $\overline{\theta}$ is estimated as a function of the average of the estimated sampling variance for each repetition estimate (termed the "within" component) and a between imputation variance component that captures the imputation variability over the repetitions of the imputation process.

The within-imputation variance component is computed as the average of the estimated variances for the $\hat{\theta}_m$ from the $m$=1,…,$M$ completed data set analyses:

$$\overline{W} = \text{Within - imputation variance} = \frac{1}{M}\sum_{m=1}^{M}\hat{W}_m = \frac{1}{M}\sum_{m=1}^{M} var(\hat{\theta}_m);$$

$$\text{where : } \text{var}(\hat{\theta}_m) \text{ is the estimate of the variance of } \hat{\theta}_m \text{ for MI repetition } m = 1,...,M.$$

The between-imputation component of the MI variance is estimated using the formula:

$$B = \text{Between - imputation variance} = \frac{1}{(M-1)}\sum_{m=1}^{M}(\hat{\theta}_m - \overline{\theta})^2.$$

The total variance of the MI estimate of θ is then computed using Rubin's combining formula:

$$\text{var}(\overline{\theta}) = \overline{W} + \left(\frac{M+1}{M}\right) \times B$$

In applications to complex sample survey data, the analysis step to generate the repetition estimates and standard errors is performed using SAS SURVEY procedures. The estimated sampling variance of each repetition estimate, $\text{var}(\hat{\theta}_m)$, is computed using an appropriate Taylor Series or replication variance estimator.

### 2.5.2 Multiple Imputation—Confidence Intervals

Based on the MI estimates of the parameter of interest and its variance, Rubin (1987) shows that the statistic:

$$t = \frac{(\theta - \overline{\theta})}{\sqrt{T}}$$
$$\text{where : } T = \text{var}(\overline{\theta})$$

is approximately distributed as a Student *t* with degrees of freedom equal to:

$$v_{mi} = (M-1)\left\{1 + r^{-1}\right\}^2$$
$$\text{where :}$$
$$r = \frac{(1 + M^{-1}) \times B}{\overline{W}}$$

The factor, *r*, in this expression is labeled the "increase in variance due to nonresponse." When there is no missing data for the variables required to estimate *θ*, both B, the between-imputation variance, and *r* will be zero.

A related measure of the impact of missing data on the variance of parameter estimates is λ, or the "fraction of missing information (FMI)" about *θ*. No matter how good the imputation model and methods are, the imputed data set will never achieve the same level of "statistical information" that would have existed in the complete, fully observed data. Based on the estimates of the within and between components of the multiple imputation variance estimator, the "fraction of missing information" is computed in SAS as:

$$\hat{\lambda} = \frac{r + 2(v_{mi} + 3)}{r + 1}$$

This statistic measures the proportion of information lost due to imputation relative to the full information that would be present if all data values for all cases had actually been observed. Since the fraction of missing information is a function of the within and between variance of specific estimates, it is specific to that estimate. In a single regression model, the fraction of missing information for one estimated parameter, $\hat{\beta}_j$, may differ from that for another, $\hat{\beta}_k$. If missing data for a single variable, *y*, were imputed based only on the distribution of the observed values of that same variable, the fraction of missing information for the estimated mean, $\overline{y}$, would equal the missing data rate for *y*. However, more generally, when the model for imputing for *y* conditions on observed values of other related variables, the imputation borrows strength from the multivariate relationships and the fraction of information lost will be reduced such that $0 < \hat{\lambda} <$ missing data rate for y.

The unique feature of the MI confidence interval for population statistics or parameter values is the degrees of freedom determination for the Student *t* distribution. In practical problems where the degrees of freedom for a complete data analysis, $v_0$, is small and the proportion of missing data is also small, the computed MI degrees of freedom approximation, $v_{mi}$, may be greater than $v_0$—a theoretical impossibility. This situation can occur in

practice when the number of independent sample observations, *n*, is small or in complex sample data designs where the effective complete data degrees of freedom is determined by the numbers of primary strata and clusters and not simply by the number of unique data observations. To ensure that the MI degrees of freedom are correctly bounded and better approximate the true degrees of freedom for the Student *t* reference distribution, SAS incorporates the "small sample" method of Barnard and Rubin (1999) to determine the degrees of freedom for constructing the confidence interval:

$$v_{mi}^{*} = \left[ \frac{1}{v_{mi}} + \frac{1}{\hat{v}_{obs}} \right]^{-1}$$

$$where:$$
$$\hat{v}_{obs} = (1-\gamma) v_0 (v_0 + 1) / (v_0 + 3);$$
$$v_o = \text{complete data degrees of freedom};$$
$$\gamma = (1 + M^{-1}) B / T.$$

The default value used by PROC MI for the complete case degrees parameter is $v_0$=infinite. Under this default, it is clear that the Barnard-Rubin small sample approximation for degrees of freedom reduces to the original MI degrees of freedom, $v_{mi}$. However, the EDF option for the MI procedure permits the analyst to override the default and specify a finite, known value for the complete data degrees of freedom. As a case in point, the complete case degrees of freedom for data collected under a stratified, clustered complex sample design is generally approximately as $v_0$=(# PSUs – # Strata). In later chapters, examples based on the NHANES, HRS, or NCS-R survey data sets will employ the EDF option on the PROC MIANALYZE command line to set the complete case degrees of freedom to the appropriate approximation for their complex sample design (e.g., EDF=16 for NHANES 2009–2010).

Therefore, MI confidence intervals for descriptive population statistics or single parameters in models are constructed from the multiple imputation estimate, its standard error, and a critical value from the Student *t* distribution (Rubin and Schenker 1986) with $v_{mi}^{*}$ degrees of freedom:

$$CI_{(1-\alpha)}(\theta) = \overline{\theta} \pm t_{v_{mi}^{*},1-\alpha/2} \cdot \sqrt{T}$$

*where* :

$v_{mi}^{*}$ = the Barnard-Rubin estimate of the MI degrees of freedom.

MI confidence intervals are routinely reported for individual population statistics or model parameters computed by PROC MI or PROC MIANALYZE. Simulation studies have demonstrated that for large sample sizes, this MI confidence interval provides true coverage of the population value that is very close to the nominal (1-α)% coverage level.

## 2.6 MI Procedures for Multivariate Inference

### 2.6.1 Multiple Parameter Hypothesis Tests

In regression and other multivariate statistical procedures, it is common to test multiple parameter hypotheses of the form, H0: β={β1,…,βq}={0,…,0}. MI procedures for making multivariate inferences of this form (i.e., the MULT option in PROC MIANALYZE) are a direct extension of those for constructing a confidence interval or *t* test statistics for single parameters. The MI estimates for the vector of parameters is computed by averaging the *m*=1,…,*M* repetition estimates of the multiparameter vector:

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m$$

$\bar{W}$ is a $q$ x $q$ within imputation covariance matrix for the $q$ x 1 parameter vector computed by averaging the covariance matrices from the $m$=1,…,$M$ repetitions:

$$\bar{W} = \frac{1}{M}\sum_{m=1}^{M}\hat{W}_m$$

The between-imputation covariance matrix is defined as:

$$B = \left(\frac{1}{M-1}\right)\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta})(\hat{\theta}_m - \bar{\theta})'$$

T, or the total covariance matrix, is defined as:

$$T = \bar{W} + \left(1 + \tfrac{1}{M}\right)B$$

Unfortunately, in many practical applications where the number (q) of multivariate parameters in θ is large relative to the number of MI repetitions (M), estimates of the between-imputation covariance matrix B are unstable. For this reason, PROC MIANALYZE uses a more stable estimate of T:

$$\tilde{T} = (1+r)\bar{W}$$

*where :*

$$r = (1 + 1/M) \times trace(B\bar{W}^{-1}) / q$$

= the average of the relative increase in variance due to missing data for the $j = 1,\ldots,q$ parameters in $\theta$.

To test the multivariate null hypotheses, $H0: \bar{\theta} = 0$, a special Wald Test $F$ statistic is used:

$$F = (\bar{\theta} - \theta_0)'\tilde{T}^{-1}(\bar{\theta} - \theta_0) / q$$

Under the null hypothesis, this test statistic is referred to a central $F$ distribution with $q$ and

$$v_1 = \frac{1}{2}(p+1)(M-1)(1 + 1/r)^2$$

degrees of freedom, where:

$$r = (1 + \frac{1}{M}) \times trace(B\bar{W}^{-1}) / q$$

For situations where $M$ is larger (i.e., $q(M-1) > 4$), PROC MIANALYZE uses an alternate expression for the denominator degrees of freedom that was proposed by Li, Raghunathan, and Rubin (1991):

$$v_2 = r + (t-4)\left[1 + \frac{1}{r} \times \left(1 - \frac{2}{t}\right)\right]^2$$

### 2.6.2 Tests of Linear Hypotheses

In the various forms of "regression analysis," linear tests of hypotheses about the parameters can be expressed in matrix notation as: $H_0 : L\beta = c$ where β is a vector of regression parameters, L is a matrix of coefficients that define the test, and $c$ is a vector of constants. In many cases, the hypothesis of interest sets $c$=0. To test such

linear hypotheses in PROC MIANALYZE, the standard TEST statement is employed with the MULT option to generate a Wald *F* test statistic to test the linear hypotheses.

## 2.7 How Many Multiple Imputation Repetitions Are Needed?

Theoretically, the statistical efficiency of multiple imputation methods is maximized when the number of repetitions is infinite, $M=\infty$. Fortunately, the same theory tells us that if we make the practical choice of using only a modest, finite number of repetitions (e.g., $M$=5, 10, or 20), that loss of efficiency compared to the theoretical maximum is relatively small. A measure of relative efficiency reported in SAS outputs from MI analysis is:

$$\mathrm{RE} = \left(1 + \frac{\lambda}{M}\right)^{-1}$$

where :
$\lambda$ is the fraction of missing information;
and $M$ is the number of MI repetitions

If the rates of missing data and therefore fraction of missing information are modest (< 20%), MI analyses based on as few as $M$=5 or $M$=10 repetitions will achieve > 96% of the maximum statistical efficiency. If the fraction of missing information is high (30% to 50%), analysts are advised to specify $M$=20 or $M$=30 to maintain a minimum relative efficiency of 95% or greater. Historically, the rule of thumb for most practical applications of MI was to use $M$=5, and this is the current default in PROC MI. Recent research has shown benefits in using larger numbers of repetitions to achieve better nominal coverage for MI confidence intervals or nominal power levels for MI hypothesis tests. Van Buuren (2012) suggests a practical "stepped" approach in which all initial MI analyses are conducted using $M$=5 repetitions. When the analyses have reached the point where a final model has been identified, the imputation can be repeated with $M$=30 or $M$=50 repetitions to ensure that the final results do not suffer from a relative efficiency loss. In offering this advice, the author also notes that this last confirmation step is unlikely to alter any conclusions that would have been drawn based on the $M$=5 repetition analyses.

## 2.8 Summary

This chapter has provided a theoretical foundation underlying the multiple imputation process. With this foundation in place, later chapters cover details of PROC MI and PROC MIANALYZE syntax along with practical applications of the MI process in SAS.



# Index

## A



## B



## C

## D



## E



## F



## G



## H



## I

## J



## K



## L



## M

## N



## O



## P

## R



## S



## T

## U



## V



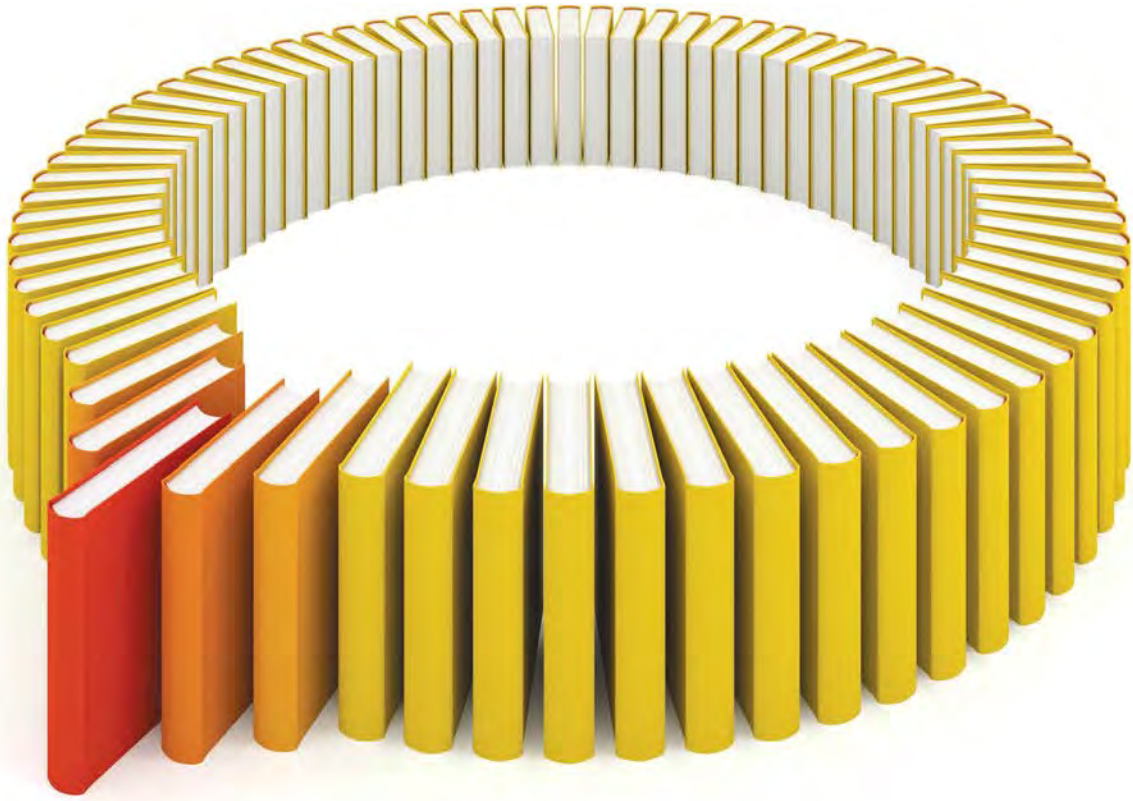
## W

# About The Authors

Patricia Berglund is a Senior Research Associate in the Survey Methodology Program at the University of Michigan Institute for Social Research (ISR). She has extensive experience in the use of SAS for data management and analysis. She is a faculty member in the ISR's Summer Institute in Survey Research Techniques and also directs the ISR's SAS training programs. Berglund also teaches a SAS Business Knowledge Series class titled "Imputation Techniques in SAS." Her primary research interests are mental health, youth substance issues, and survey methodology.

Steven Heeringa is a Senior Research Scientist at the University of Michigan Institute for Social Research (ISR) where he is Director of the Statistical Design Group. He is a member of the Faculty of the University of Michigan Program in Survey Methods and the Joint Program in Survey Methodology at the University of Maryland. Heeringa is a Fellow of the American Statistical Association and elected member of the International Statistical Institute. He is the author of many publications on statistical design and sampling methods for research in the fields of public health and the social sciences. Heeringa has over 35 years of statistical sampling experience in the development of the ISR's National Sample design, as well as research designs for the ISR's major longitudinal and cross-sectional survey programs.

Learn more about these authors by visiting their author pages, where you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more: http://support.sas.com/berglund http://support.sas.com/heeringa



# Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
*for additional books and resources.*

sas
THE POWER TO KNOW®