8

Missing data and imputation methods

Alessandra Mattei, Fabrizia Mealli and Donald B. Rubin

Missing data are a pervasive problem in many data sets and seem especially widespread in social and economic studies, such as customer satisfaction surveys. Imputation is an intuitive and flexible way to handle the incomplete data sets that result. We discuss imputation, multiple imputation (MI), and other strategies to handle missing data, together with their theoretical background. Our focus is on MI, which is a statistically valid strategy for handling missing data, although we also review other valid approaches, such as direct maximum likelihood and Bayesian methods for estimating parameters, as well as less sound methods. The creation of multiply-imputed data sets is more challenging than their analysis, but still relatively straightforward relative to other valid methods, and we discuss available software for MI. Some examples and advice on computation are provided using the ABC 2010 annual customer satisfaction survey. Ad hoc methods, including using singly-imputed data sets, almost always lead to invalid inferences and should be eschewed.

8.1 Introduction

Missing values are a common problem in many data sets and seem especially widespread in social and economic studies, including customer satisfaction surveys, where customers may fail to express their satisfaction level concerning their experience with a specific business because of lack of interest, unwillingness to criticize their sales representative, or other reasons. Unit nonresponse occurs when a selected unit (e.g., customer) does not provide any of the information being sought. Item nonresponse occurs when a unit responds to some items but not to others. Discussion of many issues related to missing data is included in the three volumes produced by the Panel on Incomplete Data of the Committee on National Statistics in 1983

Modern Analysis of Customer Surveys: with applications using R, First Edition. Edited by Ron S. Kenett and Silvia Salini.

(Madaw and Olkin, 1983; Madaw *et al.*, 1983a,b), as well as in the volume that resulted from the 1999 International Conference on Survey Nonresponse (Groves *et al.*, 2002).

Methods for analyzing incomplete data can be usefully grouped into four main categories, which are not mutually exclusive (Little and Rubin, 1987, 2002). The first group comprises procedures based on subsets of the data set without missing data, either complete-case analysis (also known as 'listwise deletion'), which discards incompletely recorded units and analyzes only the units with complete data, or available-case analysis, which discards units with incomplete data on the variables needed to calculate certain statistics. These simple methods are generally easy to use and may be satisfactory with small amounts of missing data; however, they can often lead to inefficient and biased estimates.

The second group of methods comprises weighting procedures, which deal with unit nonresponse by increasing the survey weights for responding units in the attempt to adjust for nonresponse as if it was part of the sample design. Weighting is a relatively simple alternative for reducing bias from complete-case analysis. Because these methods drop the incomplete cases, they are most useful when sampling variance is not an issue.

The third group comprises imputation-based procedures, which fill in missing values with plausible values, where the resultant completed data are then analyzed by standard methods as if there never were any missing values. In order to measure and incorporate uncertainty due to the fact that imputed values are not actual values, alternative methods have been proposed, including resampling methods and multiple imputation (MI), as proposed by Rubin (1978a, 1987, 1996). MI is a technique that replaces each set of missing values with multiple sets of plausible values representing a distribution of possibilities. Each set of imputations is used to create a complete data set, which is analyzed by complete-data methods; the results are then combined to produce estimates and confidence intervals that incorporate missing-data uncertainty. Imputation methods were originally viewed as being most appropriate in complex surveys that are used to create public-use data sets to be shared by many users, although over the years they have been successfully applied in other settings as well.

The final group of methods comprises direct analyses using model-based procedures; models are specified for the observed data, and inferences are based on likelihood or Bayesian analysis. Using these methods is typically more complex than using previous methods.

A missing-data method is required to yield statistically valid answers for scientific estimands. By a scientific estimand we mean a quantity of scientific interest that can be calculated in the population and does not change its values depending on the data collection design used to measure it. Scientific estimands include population means, variances, correlation coefficients, and regression coefficients. Inferences for a scientific estimand are defined to be statistically valid if they satisfy the following three criteria (e.g., Rubin 1996; Rässler *et al*, 2008): (a) point estimation must be approximatively unbiased for the scientific estimand; (b) interval estimation must reach at least the nominal coverage: actual interval coverage $\geq$ nominal interval coverage, so that, for example, 95% intervals for a population mean should cover the true population mean at least 95% of the time; and (c) tests of hypotheses should reject at their nominal level or less frequently when the null hypothesis is true, so that, for example, a 5% test of a zero population correlation should reject at most 5% of the time when the population correlation is zero.

In general, only MI and model-based procedures can lead to valid inferences. Resampling methods, such as the bootstrap and jackknife, can satisfy criteria (b) and (c) asymptotically, while giving no guidance on how to satisfy criterion (a) in the presence of missing data, but rather implicitly assuming that approximately unbiased estimates for scientific estimands have already been obtained; see Efron (1994) and the discussion by Rubin (1994). Such methods do not represent a complete approach to the problem of missing data, and therefore we do not discuss them further here.

This chapter is organized as follows. First, we start with a basic discussion of missing-data patterns, which describe which values are observed in the data matrix and which are missing, and missing-data mechanisms, which concern the relationship between missingness and the values of variables in the data matrix. Second, we review the four classes of approaches to handling missing data briefly introduced above, with a focus on MI, which we believe is the most generally useful approach for survey data, including customer satisfaction data. Third, a simple MI analysis is conducted for the ABC ACSS data, and results are compared to those from alternative missing-data methods. We conclude the chapter with some discussion.

8.2 Missing-data patterns and missing-data mechanisms

In order to conduct statistical analyses appropriately in the presence of missing data, it is crucial to distinguish between the missing-data pattern and the missing-data mechanism. The missing-data pattern describes which values are missing and which are observed in the data matrix. The missing-data mechanism describes to what extent missingness depends on the observed and/or unobserved data values.

8.2.1 Missing-data patterns

Let $Y = \left[y_{ij}\right]$ denote the $N \times P$ rectangular matrix of complete data, with ith row $y_i = [y_{i1}, \ldots, y_{iP}]$, where y_{ij} is the value of variable Y_j for subject i. Define $R = \left[R_{ij}\right]$, the $N \times P$ observed-data indicator matrix, with $R_{ij} = 1$ if y_{ij} is observed and $R_{ij} = 0$ if y_{ij} is missing.

Figure 8.1 shows some examples of common missing-data patterns. A simple pattern is univariate missing data, where missingness is confined to a single variable (see Figure 8.1(a)). For instance, suppose we are interested in estimating the relationship between a dependent variable Y_P, such as the overall satisfaction level with ABC, and a set of covariates (independent variables) $Y_1, \ldots, Y_{P-1}$, such as the company's continent (European versus non-European) and country, company's segmentation, and age of ABC's equipment, all of which are intended to be fully observed. Although the covariates may be fully observed, the outcome Y_P for some customers may be missing, leading to univariate missingness.

Another common pattern is obtained when the single incomplete variable Y_P in Figure 8.1(a) is replaced by a set of variables $Y_{J+1}, \ldots, Y_P$, all observed or missing on the same set of units (see Figure 8.1(b)). An example of this pattern is unit nonresponse, which occurs, for instance, if a subset of ABC customers do not complete the questionnaire about their satisfaction with ABC.

Patterns (a) and (b) are special cases of (c), monotone missing data, which is a pattern of particular interest, because methods for handling it can be easier than methods for general patterns (d). Missingness in Y is monotone if the variables can be arranged so that all $Y_{J+1}, \ldots, Y_P$ are missing when Y_J is missing, for all $J = 1, \ldots, P - 1$. In other words, the first variable in Y is at least as observed as the second variable, which is at least as observed as the third variable, and so on. Such a pattern of missingness, or a close approximation to it, is not uncommon in practice. Monotone patterns often arise in repeated-measures or longitudinal data sets, because if a unit drops out of the study in one time period, then the data will typically be missing in all subsequent time periods. Sometimes a nonmonotone missing-data pattern can be made monotone, or nearly so, by reordering the variables according to their missingness rates.

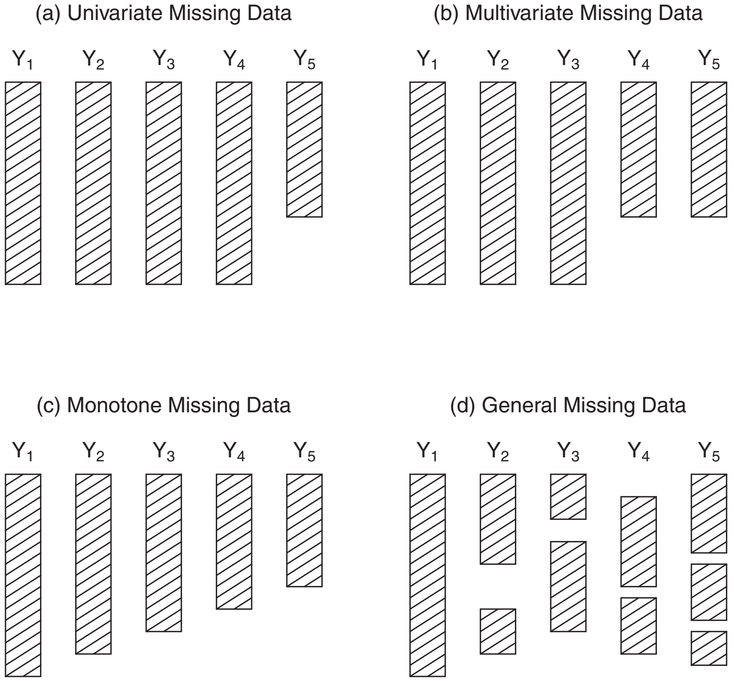
(a) Univariate Missing Data
Y1 Y2 Y3 Y4 Y5
(b) Multivariate Missing Data
Y1 Y2 Y3 Y4 Y5
(c) Monotone Missing Data
Y1 Y2 Y3 Y4 Y5
(d) General Missing Data
Y1 Y2 Y3 Y4 Y5

Figure 8.1 Examples of missing-data patterns. Rows correspond to units, columns to variables (adapted from Little and Rubin 2002, p. 5)

Sorting rows and columns of the data matrix according to the missing data to see if a simple pattern emerges may be useful. Some methods of analysis are intended for particular patterns of missing data and use only standard complete-data analyses. Other methods may be applied to more general missing-data patterns, but usually require more computational effort than methods designed for special patterns.

8.2.2 Missing-data mechanisms and ignorability

A key component in a statistical analysis with missing data is the mechanism that leads to missing data: the process that determines which values are observed, and which are missing. Missing-data mechanisms are crucial because the properties of missing-data methods strongly depend on the nature of the dependencies in these mechanisms. The missing-data mechanism is characterized by the conditional probability of the indicator matrix, R, given Y and the unknown parameters governing this process, ξ: $p\,(R|Y,\xi)$.

Following Rubin (1976), who formalized the key concepts about missing-data mechanisms, the statistical literature (e.g., Little and Rubin 2002, p. 12) classifies missing-data mechanisms into three groups: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). This language was chosen to be consistent with much older terminology in classical experimental design for completely randomized, randomized, and nonrandomized studies.

Missing data are said to be MCAR if missingness does not depend on the values of the data Y, missing or observed, that is, if the distribution of R is free of Y: $p\,(R|Y, \xi) = p\,(R|\xi)$ for all Y and ξ. In other words, missingness is MCAR if the probability that units provide data on a particular variable does not depend on the value of the variable or the value of any other variable. The MCAR assumption can be unrealistically restrictive, because it has testable implications and can be rejected by the observed data. For instance, in consumer satisfaction surveys, the MCAR assumption is contradicted by the data when companies working in different observed business areas are observed to have different rates of missing data on satisfaction.

It is generally more realistic to assume that missingness depends on observed values. For instance, the probability of missingness for satisfaction variables may depend only on completely observed variables, such as company's country, business area, age, legal status, or size, but not on any missing values. In such a case, the missing data are MAR, but not necessarily MCAR. MAR refers to missing data for which missingness can be explained by the observed values in the data set. Formally, let Y_{obs} denote the observed components of Y, and Y_{mis} the missing components. Missing data are MAR if the distribution of R depends only on Y_{obs}, and not on Y_{mis}: $p\,(R|Y, \xi) = p\,(R|Y_{\text{obs}}, \xi)$ for all Y_{mis} and ξ.

The missing-data mechanism is MNAR if the conditional distribution of R depends on the missing values in the data matrix Y, even given Y_{obs}. This could be the case with ABC if customers with lower satisfaction levels tend to be less likely to provide their satisfaction level than customers with higher overall satisfaction levels, even though they have exactly the same observed values of background covariates. The richer the data set in terms of observed variables, the more plausible the MAR assumption is.

In addition to formally defining the concepts underlying MCAR, MAR and MNAR, Rubin (1976) defined the concept of ignorability. Suppose that parametric models have been specified for both the distribution of the complete data, $p\,(Y|\psi)$, and the missing-data mechanism, $p\,(R|Y, \xi)$. The missing-data mechanism is ignorable for likelihood or Bayesian inference if: (i) the missing data are MAR; and (ii) the parameters of the data distribution, ψ, and the missing-data mechanism, ξ, are distinct, that is, the joint parameter space of (ψ, ξ) factorizes into the product of the parameter space of ψ and the parameter space of ξ, and when prior distributions are specified for ψ and ξ in a Bayesian setting, these are independent.

Ignorable missing-data mechanisms are desirable, because they allow us to obtain valid inferences about the estimands of interest, ignoring the process that causes missing data. Formally, the distribution of the observed data is obtained by integrating Y_{mis} out of the joint density of $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ and R; and the full likelihood function of ψ and ξ is proportional to it:

$$\mathcal{L}_{\text{full}}\,(\psi, \xi|Y_{\text{obs}}, R) \propto p\,(Y_{\text{obs}}, R|\psi, \xi) = \int p\,(Y_{\text{obs}}, Y_{\text{mis}}|\psi)\, p\,(R|Y_{\text{obs}}, Y_{\text{mis}}, \xi)\, \mathrm{d}\,Y_{\text{mis}}. \quad (8.1)$$

Rubin (1976) showed that if the missing-data mechanism is ignorable (MAR and distinctness of ψ and ξ), then the full likelihood function (8.1) is proportional to a simpler likelihood function,

$$\mathcal{L}_{\text{ign}}\,(\psi|Y_{\text{obs}}) \propto \int p\,(Y_{\text{obs}}, Y_{\text{mis}}|\psi)\, \mathrm{d}\,Y_{\text{mis}}, \quad (8.2)$$

which does not depend on the missing-data mechanism. Therefore valid inference about the distribution of the data can be obtained using this simpler likelihood function, ignoring the missing-data mechanism.

MAR is typically regarded as the more important condition in considerations of ignorability, because if the missing data are MAR but distinctness does not hold, inferences based on the likelihood ignoring the missing-data mechanism are still potentially valid in the sense of satisfying criteria (a)–(c) of Section 8.1, but may not be fully efficient; see (Little and Rubin 2002, Section 6.2) and Rubin (1978b) for further discussion of these ideas. Also, in many cases, it is reasonable to assume that the parameters of the data distribution and the missing-data mechanism are distinct, so that the practical question of whether the missing-data mechanism is ignorable often reduces to a question about whether the MAR assumption is plausible. This argument requires some care, however, with random effects models, where there is a subtle interplay between the assumptions of MAR and distinctness, depending on the definition of the hypothetical complete-data (see Shih 1992).

In many missing-data contexts, it is not known whether or not the ignorability condition is correct; however, assuming it can be advantageous for a variety of reasons. First, ignorability can simplify analyses greatly. Second, the MAR assumption is often reasonable, especially when there are fully observed covariates available in the data set to ‘explain’ the reasons for the missingness. Unfortunately, data can never provide any direct evidence against MAR, so that MAR is not testable without auxiliary information, such as distributional assumptions; see the literature on selection models, for example, Heckman (1976) and Little (1985). Third, even if the missing data are MNAR, an analysis based on the MAR assumption can be helpful in reducing bias by effectively imputing missing data using relationships that are observed. Finally, if the missing data are MNAR, it is usually not at all easy to specify a plausible nonignorable missing-data model, because there is no direct evidence concerning the relationship of missingness to the missing values, since the missing values are, by definition, not observed (e.g., Rubin *et al.*, 1995). Moreover, ignorable models can lead to superior inferences than misspecified nonignorable models. In addition, even if the missing-data mechanism is correctly specified, information for estimating ψ and ξ jointly may be very limited, resting strongly on the untestable assumptions made about the distribution of Y.

8.3 Simple approaches to the missing-data problem

8.3.1 Complete-case analysis

A very simple approach to missing data is to exclude incomplete units, and to use only units with all variables observed. This means that all units (cases) with any missing variables are discarded, and complete-case analysis is carried out. Complete-case analysis (sometimes called listwise deletion) is the simplest approach to missing data, because standard complete-data statistical analysis can be directly applied without modification. In addition, it guarantees comparability of univariate statistics, since they are all calculated on a common sample of cases. However, complete-case analysis may have serious pitfalls, stemming from the potential loss of information in discarding incomplete cases. Specifically, the complete-case approach leads to a loss of precision and is generally biased when the missing-data mechanism is not MCAR. The degree of bias and loss of precision depends on (i) the amount and pattern of missing data; (ii) the degree to which the MCAR assumption is violated; and (iii) the estimand and the analysis being implemented.

Complete-case analysis is potentially wasteful for univariate analysis, because values of a particular variable are discarded when they belong to cases that are missing other variables. As a result, even when complete-case analysis is unbiased, it can be highly inefficient, especially with highly multivariate data sets. For example, consider a data set with 10 variables, each of which has probability of being missing of 0.1, and suppose that missingness on each variable is independent of missingness on the other variables. Then, the expected proportion of complete cases is $(1 - 0.1)^{10} = 0.35$, that is, the complete-case analysis would be expected to include only 35% of the units.

8.3.2 Available-case analysis

Another simple approach to missing data is available-case analysis, which uses only units with complete data on the variables that are needed for the analysis being considered. This approach can be regarded as complete-case analysis restricted to the variables of interest. Available-case analysis also arises when any variable with missing values is excluded from the analysis (sometimes called 'complete-variables analysis'). Available-case analysis retains at least as many of the data values as does complete-case analysis. A drawback of this approach is that the sample base generally changes from analysis to analysis. This variability in the sample base may be problematic, because different analyses will be based on different subsets of the data and thus will not necessarily be consistent with each other. For instance, when tables are computed for various conceptual sample bases (e.g., all customers, European customers, customers working in a specific business area), the changes in the sample bases in available-case analysis prevent associating a fixed sample size to each base. These changes in the sample bases also yield problems of comparability across analyses if the missing-data mechanism is not MCAR, and may lead to misleading results when estimates of quantities concerning different variables are combined. For instance, if summaries of different variables are to be compared, the set of units for which each variable is summarized can differ across variables, and the summaries can be incomparable if the assumption of MCAR is violated. As an extreme example in the context of combining estimates, consider the estimation of the covariance of two variables and their standard deviations using available-case analysis independently for each of the three statistics; when these estimates are combined to estimate the correlation between the two variables, the resulting estimated correlation can lie outside the range $[-1, 1]$.

Complete-case analysis and available-case analysis (and combinations thereof) are extremely common approaches to handling missing data, and either was often the default strategy for handling incomplete data in older software packages. Although they are simple to implement, which is undeniably seductive, they can have serious deficiencies, which can be avoided using more modern and appropriate methods.

8.3.3 Weighting adjustment for unit nonresponse

A relatively simple device for removing or reducing the bias from complete-case analysis when the missing-data mechanism is not MCAR is to assign a nonresponse weight to each complete case (i.e., each respondent). In probability sampling, sampled units are often weighted by the inverse of their probabilities of selection in order to adjust estimates of population quantities for differential selection probabilities. The basic idea underlying weighting adjustments is to treat the complete cases as an extra layer of selection, and then to weight each complete case by the product of the sampling weight and the inverse of the conditional probability of being a complete case given selection into the sample. Although sampling weights are determined by the sample design and hence are known, nonresponse weights are based on unknown nonresponse probabilities, which need to be estimated from the data.

Nonresponse weights are generally based on background information that is available for all of the units in the survey. For instance, when a nonrespondent matches a respondent with respect to background variables measured for both, the nonrespondent's weight is simply added to the matching respondent's weight, and the nonrespondent is discarded. Because the match is defined by observed variables, such adjustment implicitly assumes MAR: if the MAR assumption is satisfied, weighting, in principle, removes nonresponse bias. In order to increase the plausibility of the MAR assumption, it is important to attempt to record background characteristics of respondents and nonrespondents that are predictive of nonresponse and use these variables to define nonresponse weights. Background characteristics should also be predictive of survey outcomes to limit sampling variance of resulting estimates.

Weighting methods can be useful for removing or reducing the bias in complete-case analysis. However, weighting methods do have some serious pitfalls. First, weighted estimates can have unacceptably high sampling variance due to the possibility of large weights. Second, the computation of appropriate standard errors for weighted complete-case estimators is often problematic. Explicit formulas are available for simple estimators, but methods are not well developed for more complex situations. Finally, the use of such weighting adjustment when dealing with item nonresponse deletes all incomplete cases and so discards additional observed data, which are not used in creating the weighting adjustment. For further discussion of weighting procedures for nonresponse in general, see Bethlehem (2002), Gelman and Carlin (2002), Little and Schenker (1995), and Little and Rubin (2002, Section 3.3).

8.4 Single imputation

Both complete-case and available-case analysis generally discard units with some observed data. An attractive alternative approach for handling incomplete data is to impute (fill in) the values of the items that are missing. A variety of imputation approaches can be used that range from extremely simple to rather complex. These methods can be applied to impute one value for each missing item (single imputation) or, in some situations, to impute more than one value, to allow appropriate assessment of imputation uncertainty (multiple imputation). Imputations are typically created assuming that the missing-data mechanism is ignorable, and for simplicity, here we focus our discussion on the ignorable situation.

Good imputations are draws from the predictive distribution of the missing values. A method for creating a predictive distribution for the imputations based on the observed data is required. This distribution can be generated by using either an explicit or an implicit modeling approach or a combination of the two approaches. The first approach requires the specification of a formal statistical model, on which the predictive distribution is based with explicitly stated assumptions. In the implicit modeling approach, the focus is on an algorithm, which implies an underlying model. Although assumptions are now implicit, they still need to be carefully assessed to ensure that they are reasonable.

Explicit modeling methods include: *mean imputation*, where missing values are replaced by means from the responding units in the sample; *regression imputation*, which replaces missing values by predicted values from a regression of the missing item on items observed for the unit, usually calculated from units with both sets of variables present; and *stochastic* *regression imputation*, which replaces missing values by a value predicted by regression imputation plus a residual drawn to reflect uncertainty in the predicted value.

In survey practice a common implicit modeling method is *hot-deck imputation*. Hot-deck imputation replaces each missing value with a random draw from a 'donor pool' consisting of values of that variable observed on responding units similar to the unit with the missing value. A donor pool is selected, for instance, by choosing units with complete data who have 'similar' observed values to the unit with missing data, for example, by exact matching on their observed values, or using a distance measure (metric) on observed variables to define 'similar'.

Singly imputed data sets are straightforward to analyze using standard complete-data methods; however, creating decent imputations may require substantial effort. Little and Rubin (2002) suggest some guidelines for creating imputations. Specifically, imputations should be: (1) conditional on observed variables, to reduce bias due to nonresponse, improve precision, and reflect associations between missing and observed variables; (2) multivariate, to preserve associations between missing variables; and (3) randomly drawn from predictive distributions, rather than set equal to means, to account properly for variability.

Unconditional mean imputation, which replaces each missing value with the mean of the observed values of that variable, meets none of the three guidelines. Conditional mean imputation, which replaces missing values of each variable with the mean of that variable calculated within cells defined by observed categorical variables, and regression imputation can satisfy the first two guidelines. Only stochastic regression imputation and hot-deck imputation, when done properly, can meet all three guidelines for single imputation.

Singly imputed data sets, created following the three guidelines suggested by Little and Rubin (2002), can be analyzed using standard complete-data techniques. The resulting inferences can satisfy criterion (a) of Section 8.1, leading to approximately unbiased estimates under ignorability. However, such inferences nearly always fail to satisfy criteria (b) and (c), providing too small estimated standard errors, too narrow confidence intervals, and too significant p-values for hypothesis tests, regardless of how the imputations were created. The reason is that the automatic application of standard complete-data methods to singly-imputed data sets treats imputed values as if they were known, although they are actually not known. In other words, single imputation followed by a complete-data analysis that does not distinguish between real and imputed values is almost always statistically invalid, because inferences about estimands based on the filled-in data do not account for imputation uncertainty.

Special methods for sampling variance estimation following single imputation have been developed for specific imputation procedures and estimation problems (e.g., Schafer and Schenker 2000; and Lee *et al.*, 2002). However, such techniques need to be customized to the imputation method used and to the analysis method at hand, and they often require the user to have information about the imputation model that is not typically available in shared data sets.

As an alternative, multiple imputation can be carried out. Multiple imputation is less computationally intensive than the replication approach (e.g., Efron 1994; Shao 2002), and generally can lead to valid inferences in the sense of satisfying criteria (a)–(c) of Section 8.1. Multiple imputation accounts for missing data by not only restoring the natural variability in the missing data, but also incorporating the uncertainty created by predicting missing data. Maintaining the original variability of the missing data is done by creating imputed values that are based on variables correlated with the missing data and reasons for the missingness.

Uncertainty is accounted for by creating different versions of the missing data and using the variability between imputed data sets.

8.5 Multiple imputation

Multiple imputation was first proposed in Rubin (1977, 1978a) and discussed in detail in Rubin (1987, 1996, 2004a,b). MI is a simulation technique that replaces each missing value in Y_{mis} with a vector of $m > 1$ plausible imputed values. These m values are ordered in the sense that m completed data sets can be created by the set of vectors of imputations; replacing each missing value by the first component in its vector of imputation creates the first complete data set, replacing each missing value by the second component in its vector of imputation creates the second complete data set, and so on. Thus, m completed data sets are created: $Y^{(1)}, \ldots, Y^{(\ell)}, \ldots, Y^{(m)}$, where $Y^{\ell} = \left(Y_{\text{obs}}, Y_{\text{mis}}^{(\ell)}\right)$. Typically m is fairly small: $m = 5$ is a standard number of imputations to use. Rubin (1987) showed that the relative efficiency of an estimate based on m imputations to one based on an infinite number of them is approximately $(1 + \gamma_0/m)^{-1/2}$ in units of standard errors, where γ_0 is the population fraction of missing information.[1] Therefore, unless rates of missing information are unusually high, there is often limited practical benefit to using more than five to ten imputations, except when conducting multi-component tests.

MI retains the advantages of single imputation while allowing the uncertainty due to the process of imputation to be directly assessed and included to create statistically valid inferences. Specifically, each of the m completed data sets is analyzed using standard complete-data procedures. When the m sets of imputations are repeated random draws from the predictive distribution of the missing values under a particular missing-data mechanism, the m complete-data inferences can be easily combined to form one inference that appropriately reflects both sampling variability and missing-data uncertainty. When the imputations are from two or more models for nonresponse, the combined inferences under the models can be contrasted across models to display the sensitivity of inference to alternative missing-data mechanisms.

Most of the techniques presently available for creating MIs assume that the missing-data mechanism is ignorable, but it is important to note that the MI paradigm does not require ignorable nonresponse. MIs may, in principle, be created under any kind of model for the missing-data mechanism, and the resulting inferences will be valid under that mechanism (see Rubin 1987, Chapter 6). Schafer (1997) is an excellent source of computational guidance for creating multiple imputations under a variety of models for the data.

8.5.1 Multiple-imputation inference for a scalar estimand

The analysis of a multiply imputed data set is quite direct. First, each data set completed by imputation is analyzed using the standard complete-data method that would be used in the absence of nonresponse. Let θ be the scalar estimand of interest (e.g., the mean of a variable, or the proportion of customers who are highly satisfied with a service). Let $\widehat{\theta}$ and $\widehat{V}$ be the complete-data estimators of θ and the sampling variance of $\widehat{\theta}$, respectively. Also, let $\widehat{\theta}_{\ell}$ and $\widehat{V}_{\ell}$, $\ell = 1, \ldots, m$,

[1] In the simple case of univariate missingness and no covariates, γ_0 is equal to the expected fraction of missing units (see Section 8.5.1). When there are many variables in a survey, however, γ_0 is typically smaller than this fraction because of the dependence between variables and the resulting ability to improve prediction of missing values from observed ones.

be m complete-data estimates of θ and their associated sampling variances, calculated from m repeated imputations under one missing-data model. The m sets of statistics are combined to produce the final point estimate: $\widehat{\theta}_{\text{MI}} = m^{-1}\sum_{\ell=1}^{m}\widehat{\theta}_{\ell}$. The variability associated with this estimate has two components: the average within-imputation variance $W_m = m^{-1}\sum_{\ell=1}^{m}\widehat{V}_{\ell}$, and the between-imputation variance $B_m = (m-1)^{-1}\sum_{\ell=1}^{m}\left(\widehat{\theta}_{\ell} - \widehat{\theta}_{\text{MI}}\right)^2$. The total variability associated with $\widehat{\theta}_{\text{MI}}$ is $T_m = W_m + \left(1 + m^{-1}\right)B_m$, where the factor $1 + m^{-1}$ reflects the fact that only a finite number of completed-data estimates $\widehat{\theta}_{\ell}$, $\ell = 1, \ldots, m$, are averaged together to obtain the final point estimate (Rubin 1987, pp. 87–94).

The reference distribution for interval estimates and significance tests for θ is a Student t distribution: $\left(\theta - \widehat{\theta}_{\text{MI}}\right)T_m^{-1/2} \sim t$. Under the assumption that, with complete data, a normal reference distribution would be appropriate, the degrees of freedom of the t distribution can be approximated by the value $\nu = (m-1)(1 + r_m^{-1})^2$, where $r_m = \left(1 + m^{-1}\right)B_m/W_m$ is the relative increase in variance due to nonresponse (Rubin 1987; Rubin and Schenker 1986). Barnard and Rubin (1999) relaxed the normality assumption to allow Student t reference distributions with complete data, and proposed the small-sample adjusted value $\nu_{\text{BR}} = \left(\nu^{-1} + \nu_{\text{obs}}^{-1}\right)^{-1}$ for the degrees of freedom of the t distribution in the MI analysis, where $\nu_{\text{obs}} = (1 + r_m)^{-1}\nu_{\text{com}}(\nu_{\text{com}} + 1)(\nu_{\text{com}} + 3)^{-1}$, and ν_{com} is the complete-data degrees of freedom. Another useful statistic about the nonresponse is the fraction of missing information due to nonresponse: $\gamma_m = (1 - (\nu + 1)/(\nu + 3)W_m/t_m)$, or the generalization proposed by Barnard and Rubin (1999) $\gamma_m = (1 - \nu_{\text{BR}} + 1)/(\nu_{\text{BR}} + 3)W_m/T_m)(\nu_{\text{com}} + 3)/(\nu_{\text{com}} + 1)$.

Inferential questions that cannot be cast in terms of a one-dimensional estimand can be handled through multivariate generalizations of this rule. See, for instance, Li *et al.* (1991a, 1991b), Rubin and Schenker (1991), Meng and Rubin (1992), and Little and Rubin (2002, Section 10.2) for additional methods for combining vector-valued estimates, significance levels, and likelihood ratio statistics.

8.5.2 Proper multiple imputation

The great virtues of MI are its simplicity and its generality. The user may analyze the data by virtually any technique that would be appropriate if the data were complete. The validity of the method, however, hinges on how the imputations $Y_{\text{mis}}^{(1)}, \ldots, Y_{\text{mis}}^{(m)}$ are generated. Clearly it is not possible to obtain valid inferences in general if imputations are created arbitrarily. The imputations should, on average, give reasonable predictions for the missing data, and the variability among them must reflect an appropriate degree of uncertainty. Rubin (1987) provides technical conditions under which repeated-imputation methods lead to statistically valid answers. An imputation method that satisfies these conditions is said to be 'proper'. The term 'proper' basically means that the summary statistics $\widehat{\theta}_{\text{MI}}$, W_m and B_m, previously defined, yield approximately valid inference for the complete-data statistics, $\widehat{\theta}$ and $\widehat{V}$, over repeated realizations of the missing-data mechanism. Specifically, a multiple-imputation procedure is proper for the complete-data statistics, $\widehat{\theta}$ and $\widehat{V}$, if the following three conditions are satisfied: (1) as $m \to \infty$, $\left(\widehat{\theta}_{\text{MI}} - \widehat{\theta}\right)/\sqrt{B_m}$ converges in distribution to a $\mathcal{N}(0, 1)$ random variable over the distribution of the response indicators R with Y held fixed; (2) W_m is a consistent estimate of $\widehat{V}$ as $m \to \infty$, with R regarded as random and Y regarded as fixed; (3) treating Y as fixed, the variability of the variance of $\widehat{\theta}_{\text{MI}}$ over an infinite number of multiple imputations is of lower order than that of $\widehat{\theta}$.

These conditions are useful for evaluating the properties of an imputation method but provide little guidance for one seeking to create such a method in practice. For this reason, it is recommended that imputations be created through Bayesian arguments. For notational simplicity, assume that the missing-data mechanism is ignorable. Proper imputations are often most easily obtained as independent random draws from the posterior predictive distribution of the missing data given the observed data. Given a parametric model for the complete data, $p\,(Y_{\text{obs}}, Y_{\text{mis}}|\psi)$, and a prior distribution for the unknown model parameters, $p(\psi)$, the posterior predictive distribution of Y_{mis} given Y_{obs} can be formally written as $p\,(Y_{\text{mis}}|Y_{\text{obs}}) = \int p\,(Y_{\text{mis}}, \psi|Y_{\text{obs}})\,\mathrm{d}\,\psi = \int p\,(Y_{\text{mis}}|Y_{\text{obs}}, \psi)\,p\,(\psi|Y_{\text{obs}})\,\mathrm{d}\,\psi$. The distribution $p\,(Y_{\text{mis}}|Y_{\text{obs}})$ is a 'posterior' distribution because it is conditional on the observed data, Y_{obs}, and it is a 'predictive' distribution because it predicts the missing data, Y_{mis}. Imputations crated as independent realizations from $p\,(Y_{\text{mis}}|Y_{\text{obs}})$ can be proper because they reflect uncertainty about Y_{mis} given the parameters of the complete-data model, as well as uncertainty about the unknown model parameters, by taking draws of ψ from its posterior distribution, $p\,(\psi|Y_{\text{obs}})$, before using ψ to impute the missing data, Y_{mis}, from $p\,(Y_{\text{mis}}|Y_{\text{obs}}, \psi)$.

Imputations methods that do not account for all sources of variability are defined to be improper by Rubin (1987, Chapter 4). Thus, for instance, fixing ψ at a point estimate $\widehat{\psi}$ and then drawing m imputations for Y_{mis} independently from $p\left(Y_{\text{mis}}|Y_{\text{obs}}, \widehat{\psi}\right)$ would constitute an improper MI procedure.

8.5.3 Appropriately drawing imputations with monotone missing-data patterns

When there are many variables to be imputed, drawing random samples from the posterior predictive distribution, $p\,(Y_{\text{mis}}|Y_{\text{obs}})$, may be difficult and require high-level expertise in both statistical computing methodology and software development. In a principled modeling approach, filling in the entire set of missing data, Y_{mis}, requires postulating a joint model for all variables with any missingness given the other variables, which has to be flexible enough to reflect the structure of complex data, which may include continuous, semicontinuous, ordinal, binary, and categorical variables. However, when the missing-data pattern is monotone, creating multiple imputations is relatively straightforward because the joint distribution of all variables can be specified sequentially as the product of conditional distributions.

Specifically, suppose that $(Y_{\text{obs}}, Y_{\text{mis}}) = \left(Y^*_{\text{obs}}, Y_{\text{mis},1}, Y_{\text{mis},2}, \ldots, Y_{\text{mis},k}\right)$ follows a monotone pattern of missingness, where Y^*_{obs} represents the fully observed variables, and $Y_{\text{mis},1}$ is the incompletely observed variable with the fewest missing values, $Y_{\text{mis},2}$ the variable with the second fewest missing values, and so on, $Y_{\text{mis},k}$ being the incompletely observed variable with the most missing values. Proper imputation with a monotone missing-data pattern begins by fitting an appropriate model to predict $Y_{\text{mis},1}$ from Y^*_{obs} and then using this model to impute the missing values in $Y_{\text{mis},1}$. For example, a regression model of $Y_{\text{mis},1}$ on Y^*_{obs} can be fitted using units with $Y_{\text{mis},1}$ observed, then the regression parameters of this model are drawn from their posterior distribution, and the missing values of $Y_{\text{mis},1}$ are drawn from the posterior distribution of $Y_{\text{mis},1}$ given these drawn parameters and the observed values of Y^*_{obs}. Next, the missing values for $Y_{\text{mis},2}$ are imputed using Y^*_{obs} and the observed and imputed values of $Y_{\text{mis},1}$; for example, if $Y_{\text{mis},2}$ is a binary variable, a logistic regression model for $Y_{\text{mis},2}$ given $\left(Y^*_{\text{obs}}, Y_{\text{mis},1}\right)$ could be used. Continue to impute the next most complete variable until all missing values have been imputed. In the case of monotone missing-data patterns, the product of the univariate prediction models defines the implied full imputation model, $p\,(Y_{\text{mis}}|Y_{\text{obs}})$, and the collection of imputed values is a proper imputation of the missing data, Y_{mis}, under this model.

8.5.4 Appropriately drawing imputations with nonmonotone missing-data patterns

When missingness is not monotone, creating imputations generally involves applying iterative simulation techniques, because directly drawing from $p\,(Y_{\text{mis}}|Y_{\text{obs}})$ is generally intractable. In this case, Markov chain Monte Carlo (MCMC) provides a flexible set of tools for creating MIs from parametric models. Schafer (1997) describes MCMC methods to multiply-impute rectangular data sets with arbitrary patterns of missing values when the missing-data mechanism is ignorable, and also provides data examples and practical advice. These methods are applicable when the rows of the complete-data matrix can be modeled as independent and identically distributed observations from the following multivariate models: multivariate normal for continuous data, multinomial (including log-linear models) for categorical data, and the general location model for mixed multivariate data.

One MCMC method well suited to missing-data problems is the data augmentation (DA) algorithm of Tanner and Wong (1987). Briefly, letting t index iterations, DA involves iterating between (i) randomly sampling missing data from their conditional posterior predictive distributions given current parameter values, $Y_{\text{mis}}^{(t)} \sim p\left(Y_{\text{mis}}|Y_{\text{obs}}, \psi^{(t-1)}\right)$, where $\psi^{(t-1)}$ is the current draw of unknown parameters; and (ii) randomly sampling unknown parameters from a simulated current complete-data posterior distribution, $\psi^{(t)} \sim p\left(\psi|Y_{\text{obs}}, Y_{\text{mis}}^{(t)}\right)$. Given an initial value for ψ, say $\psi^{(0)}$, this algorithm defines a Markov chain $\left\{\left(Y_{\text{mis}}^{(t)}, \psi^{(t)}\right), t = 1, 2, \ldots\right\}$, which, under quite general conditions, converges to the stationary distribution of interest, $p\,(Y_{\text{mis}}, \psi|Y_{\text{obs}})$.

Executing these steps until the Markov chain has reached effective convergence produces a draw of ψ from its observed data posterior distribution, $p\,(\psi|Y_{\text{obs}})$, and a draw of Y_{mis} from $p\,(Y_{\text{mis}}|Y_{\text{obs}})$, the distribution from which MIs are to be generated. In many cases, the second step of the algorithm, $\psi^{(t)} \sim p\left(\psi|Y_{\text{obs}}, Y_{\text{mis}}^{(t)}\right)$, is straightforward. In more complicated situations, this step is intractable and may be replaced by one or more cycles of another MCMC algorithm that converges to $p\left(\psi|Y_{\text{obs}}, Y_{\text{mis}}^{(t)}\right)$. Much software presently available for creating multiple imputations uses DA (or variants of DA) to fill in missing values. Other algorithms that use MCMC methods for imputing missing values include the Gibbs sampler (Geman and Geman 1984) and the Methopolis–Hastings algorithm (Hastings 1970; Metropolis *et al*., 1953; Metropolis and Ulam 1949). See also Gelman *et al*. (2003) and Gilks *et al.* (1996) for more details on these algorithms, and Schafer (1997) for a complete exposition of MCMC methods in the imputation setting.

An alternative and popular approach to the creation of imputations in nonmonotone incomplete multivariate data uses the Gibbs sampler with fully conditionally specified models, where the distribution of each variable given all the other variables is the starting point. For each variable, a draw of parameters estimated using units with that variable observed is made, and then the missing data are imputed for that variable, and the procedure cycles through each variable with missing values, replacing missing values that are being conditioned on in a regression by the previously imputed values. Practical implementations of this idea include Kennickell (1991), Van Buuren and Oudshoorn K. (1999), Van Buuren and Oudshoorn C.G.M. (2000), Raghunathan *et al*. (2001), Münnich and Rässler (2005), and Van Buuren *et al*. (2006). The theoretical weakness of this approach is that the specified conditional densities may be incompatible, in the sense that they cannot be derived from a single joint distribution, and therefore the stationary distribution to which the Gibbs sampler attempts to converge may not exist.

In order to minimize or eliminate such incompatibility, Baccini *et al.* (2010) apply the ‘multiple imputation by ordered monotone blocks’ (IMB) strategy to the Anthrax Vaccine Adsorbed Trial data. This approach extends the theory for monotone patterns to arbitrary missing patterns, by breaking the problem into a collection of smaller problems where missing data do form a monotone pattern. The proposal of monotone blocks is a natural extension of using a single major monotone block (Rubin 2003). The IMB algorithm can be briefly described as follows. The variables and units in the data set are first rearranged such that the missing values not forming part of a monotone block are identified as minimal. The part that is monotone is labelled the ‘first’ monotone block. For those missing values that do not belong to the first monotone block, the process is repeated, identifying a rearrangement forming a monotone block, with the rest of the missing values being minimal. The process continues until all missing values have been identified with a monotone block. After the monotone blocks are obtained, the missing data within each block are multiply imputed. MI proceeds as follows: the missing data of all but the first monotone block are filled in with preliminary values; Bayesian sequential models are used to simulate the missing values for the first monotone block; the data imputed for the first monotone block are treated as observed and the missing values for the second monotone block are imputed, again using Bayesian sequential models. This process is performed across all the monotone blocks, and iterated until apparent convergence.

8.5.5 Multiple imputation in practice

A key feature of either single or multiple imputation is that the imputation phase is operationally distinct from subsequent analysis phases. As a result, imputations may be created by one person or organization and the ultimate analyses carried out by another, and the implicit or explicit model used for creating imputations may differ from the implicit or explicit model used in subsequent analyses of the completed data. In many cases, imputations are created just once by an expert in missing-data techniques (e.g., the data collector), who may have detailed knowledge or even additional confidential data that cannot be made available to the ultimate analysts but which may be relevant to the prediction of missing values. The ultimate user of multiply-imputed data could apply a variety of potentially complicated complete-data analyses, and then use the combining rules and combined results even though the multiple imputations were created under different models.

This feature gives MI great inherent flexibility and it is especially attractive in the context of public-use data sets that are shared by many ultimate users, but raises the possibility that the statistical model or assumptions used to create the imputed data sets may be incompatible with those used to analyze them. Meng (1994) defines the imputer’s and analyst’s models as ‘congenial’ if the resulting inference is fully valid. When the imputer makes fewer assumptions than the analyst, then MI generally leads to valid inferences with perhaps some loss of efficiency, because the additional generality of the imputation model may increase variability among the imputed data sets. When the imputer makes more assumptions than the analyst – and the extra assumptions are true – then imputations may turn out ‘superefficient’ from the perspective of the data analyst (Rubin 1996), in the sense that MI inferences may be more precise than any inference derived from the observed data and the analyst’s model alone, because they reflect the imputer’s better knowledge about the process that creates nonresponse. The only serious negative effect of inconsistency arises when the imputer makes more assumptions than the analyst and these additional assumptions are false, because the multiple imputations created under an incorrect model can lead to erroneous conclusions. For further discussions on the validity of multiple-imputation inference when the imputer's and analyst's models differ, see Fay (1992), Meng (1994), and Rubin (1996).

Clearly, congeniality is more easily satisfied when the imputer and the analyst are the same entity or communicate with each other. In the context of shared data sets, however, in order to warrant near-congeniality of the imputer's and user's models, the imputation model should include a set of variables as rich as possible (e.g, Rubin 1996). In practice, this means that an imputation model should reasonably preserve those distributional features (e.g., associations) that will be the subject of future analyses. It is especially important to include design variables, such as variables used to derive sampling weights, or the sampling weights themselves, and domain indicators when domain estimates are to be obtained by subsequent users. When such critical variables are excluded from the imputation model, point estimates, as well as sampling variance estimates based on this model, will generally be biased (e.g., Kim *et al.*, 2006).

The performance of a MI procedure depends on several factors, including the posited missing-data mechanism, the (implicit or explicit) imputation model specified for the data, and the complete-data analyses the ultimate user performs. Therefore, in order to obtain valid inference from multiply-imputed data sets, the performance of the imputation strategy should be carefully assessed (e.g., Baccini *et al.*, 2009; Schafer *et al.*, 1996; Tang *et al.*, 2005).

8.5.6 Software for multiple imputation

Many statistical software packages have built-in or add-on functions for creating multiply-imputed data sets, managing the results from each imputed data set, and combining the inferences using the method described in Section 8.5.1 and its multivariate generalizations. Joseph Schafer has produced the S-Plus libraries NORM (which is also available as stand-alone Windows package), CAT, MIX and PAN for multiply imputing normal, categorical, mixed and panel data, respectively. These libraries are freely available (see http://www.stat.psu.edu/∼jls/misoftwa.html).

Multiple Imputation by Chained Equations (MICE) is another freely-available library distributed for S-Plus, which may be downloaded from the www.multiple-imputation.com Web site. Procedures to impute missing data using MICE are also implemented in other software packages, including IVEware, the R environment, STATA, and SPSS. For instance, the *mi* (Gelman *et al.*, 2011; Su *et al.*, forthcoming) and *mice* (Van Buuren and Groothuis-Oudshoorn, forthcoming) packages implement the MICE procedure within R, and STATA provides the *ice* command to impute missing data based on the chained equation approach (Royston 2004, 2005). MICE-MI is a stand-alone version of mice, and it is downloadable from http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm. The S-Plus missing data library extends S-Plus to support model-based missing data models, by using the EM (Dempster *et al.*, 1977) and DA algorithms (Tanner and Wong, 1987).

IVEware (http://www.isr.umich.edu/src/smp/ive) by Raghunathan *et al.* (2001) is very flexible and freely available software for MI; it is an SAS version 9 callable routine built using the SAS macro language or a stand-alone executable. In addition to supporting chained equations, IVEware extends multiple imputation to support complex survey sample designs.

In SAS/STAT, multiple imputation is implemented by two procedures. The imputation step is carried out by PROC MI. Then, complete-data methods are employed using any of the SAS procedures for complete-data analysis. Finally, the results are combined using PROC MIANALYZE.

SOLAS is commercially available software designed specifically for creating and analyzing multiply-imputed data sets (http://www.statsol.ie/solas/solas.htm). SOLAS is most appropriate for data sets with a monotone or nearly monotone pattern of missing data.

Other packages that provide some support for MI are currently available. We refer to the www.multiple-imputation.com website for more information, or to Horton and Lipsitz (2001) and Horton and Kleinman (2007) for some historical perspective.

8.6 Model-based approaches to the analysis of missing data

We now describe model-based missing-data methods where an explicit model for the complete data is specified and inferences are based on the likelihood or posterior distribution under that model. In full generality, statistical models are developed by specifying $p(Y, R|\psi, \xi)$, the joint distribution of Y and R (Rubin 1976). Two classes of models have been proposed, based on alternative factorizations for this distribution. Selection models (e.g., Heckman, 1976; Little and Rubin, 2002) specify the joint distribution of Y and R as

$$p(Y, R|\psi, \xi) = p(Y|\psi)\, p(R|Y, \xi), \tag{8.3}$$

where $p(Y|\psi)$ represents the complete-data model for Y, $p(R|Y, \xi)$ represents the model for the missing-data mechanism, and ψ and ξ are unknown distinct parameters. Pattern-mixture models (e.g., Rubin, 1977, 1978a; Glynn *et al.*, 1986, 1993; Little, 1993) specify

$$p(Y, R|\phi, \pi) = p(Y|R, \phi)\, p(R|\pi) \tag{8.4}$$

where $p(Y|R, \phi)$ represents the conditional distribution of Y given the missing-data pattern R, $p(R|\pi)$ models the missing-data indicator, and ϕ and π are unknown distinct parameters. Pattern-mixture models partition the data with respect to the missingness of the variables, and the resulting marginal distribution of Y is a mixture of distributions. If R is independent of Y, that is, missingness is MCAR, then these two model forms are easily seen to be equivalent with $\psi = \phi$ and $\xi = \pi$. When the missing data are not MCAR, the two specifications generally yield different models because of the different distinctness of parameters.

Little and Rubin (2002, Chapter 15) discuss the use of selection and pattern-mixture approaches in the context of nonignorable missing-data mechanisms for different types of data. As discussed previously, a crucial point about the use of nonignorable models is that they are difficult to specify correctly, because there is no direct evidence in the data about the relationship between the missing-data mechanism and the missing values themselves. For this reason, selection models and pattern-mixture models for nonignorable missing data generally depend strongly on assumptions about specific distributions, which are not directly testable. Consequently, sensitivity to model specification is a serious scientific problem for both selection and pattern-mixture models. Thus, whenever possible, it is advisable to consider several nonignorable models, rather than to rely exclusively on one model, and to explore the sensitivity of answers to the choice of the model, using a baseline analysis under ignorability as a primary point of comparison.

As discussed in Section 8.2.2, when the missing-data mechanism is ignorable, statistically valid inferences for the parameters of the data distribution, ψ, can be based on the likelihood function for ψ ignoring the missing-data mechanism: $\mathcal{L}_{\text{ign}}(\psi|Y_{\text{obs}})$ (see equation (8.2)) (Rubin 1976). Little and Rubin (2002, Chapters 11–14) provide a complete exposition of ignorable likelihood and Bayesian methods, also describing their application to solve different analytic problems, as well as reviewing several examples where analyses of incomplete data are carried out under the assumption of an ignorable missing-data mechanism.

Once maximum likelihood (ML) estimates of parameters have been obtained, inferences can be derived by applying standard methods. In many incomplete-data problems, however, the likelihood function (8.2) is a complicated function, and explicit expressions for the ML estimates of ψ are difficult to derive. Standard numerical ML algorithms, such as the Newton–Raphson algorithm, can be applied, but other iterative procedures, exploiting the missing-data aspect of the problem, may have advantages. The best known of these algorithms is the expectation–maximization (EM) algorithm (Dempster *et al.*, 1977), which takes advantage of the facts that: (1) if ψ were known, estimation of many functions of Y_{mis} would be relatively easy; and (2) if the data were complete, computation of ML estimates would be relatively simple. Several extensions of EM have been also proposed, including ECM (Meng and Rubin 1993), ECME (Liu and Rubin 1994), AECM (Meng and van Dyk 1997), and PX-EM (Liu *et al.*, 1998). For detailed discussions of the theoretical properties of the EM algorithm, examples of its use, methods for obtaining standard errors based on the algorithm, and its extensions, see Dempster *et al.* (1977), McLachlan and Krishnan (1997), Schafer (1997), and Little and Rubin (2002, Chapters 8, 9 and 11–15).

ML techniques are most useful when sample sizes are large, because then the log-likelihood is approximately quadratic in the neighborhood of the ML estimates, and can be summarized well using the ML estimates and their asymptotic covariance matrix. In small samples, ML methods may have unsatisfactory properties because the assumption of asymptotic normality of the likelihood may be unreasonable. Thus, alternatives to ML inference may be preferable. When sample sizes are small or ML techniques are intractable, simulation methods can be used, which are often easier to implement than analytic methods. From the Bayesian perspective, the focus is on iterative simulation methods for approximating the posterior distribution of the parameters of interest, ψ.

Under the assumption of an ignorable missing-data mechanism, Bayesian inferences for ψ are based on the observed-data posterior distribution with density $p(\psi|Y_{\text{obs}}) \propto p(\psi)$ $p(Y_{\text{obs}}|\psi)$, where $p(\psi)$ is the prior density for ψ. As with ML estimation, working explicitly with this observed-data posterior distribution can be difficult. The DA algorithm (Tanner and Wong 1987), introduced in Section 8.5.4, may facilitate drawing ψ from $p(\psi|Y_{\text{obs}})$. For discussions of the theoretical properties of the DA algorithm, its extensions, and examples of the use of Bayesian iterative simulation methods, see Tanner and Wong (1987), Gelfand and Smith (1990), Schafer (1997), and Little and Rubin (2002, Chapters 10–14). Although we focus on ignorable nonresponse, the EM and DA algorithms can be also applied in the context of the nonignorable missing-data mechanism (e.g., Little and Rubin 2002, Chapter 15).

If the sample size is large, likelihood-based analyses and Bayesian analyses under diffuse prior distributions are expected to provide similar results because the likelihood dominates the prior distribution and is nearly multivariate normal. If the sample size is small, a Bayesian analysis may be preferable, because it allows us to avoid the usual assumption of asymptotic normality of the likelihood.

8.7 Addressing missing data in the ABC annual customer satisfaction survey: An example

Since 2001, ABC has conducted an annual customer satisfaction survey (ACSS) to gather information on its touch points and interactions with customers through a questionnaire consisting of 81 questions. The ABC ACSS suffers from both unit nonresponse and item nonresponse. Over the years, and across geographical areas, respondents to the ACSS questionnaire range from 10% to 80%, with a typical response rate of 45%. Here, we focus on item nonresponse, using data from the 2010 ACSS, which provides information on satisfaction levels of 266 ABC customers (see Chapter 2 for a detailed description of the ACSS and preliminary analyses of the 2010 ACSS). For simplicity, we select a subset of 12 questions, including overall satisfaction with ABC, willingness to recommend ABC to other companies, repurchasing intentions, and overall satisfaction with each of the following topics: equipment, sales support, technical support, training, supplies, software solutions, customer website, purchasing support, and contracts and pricing. Each satisfaction item is an ordinal variable with five categories ($1 =$ very low satisfaction level, ..., $5 =$ very high satisfaction level). Six customer background variables are also included: country, segmentation, age of ABC's equipment, company's profitability, customer's position in the company and customer seniority. Country and customer's position are completely observed, age of ABC's equipment and customer seniority are missing for one customer, company's profitability is missing for 25 customers, and segmentation is missing for 43 customers.

The first column of Table 8.1 presents the proportion of missing values for each satisfaction variable, which shows that missingness rates are somewhat high for some variables. The missing-data pattern is not monotone and only 67 (25%) customers provide complete data. Thus, a complete-case analysis may lead to a substantial loss of information, implying a loss of precision and potential bias. An available-case analysis may be a simple alternative using more information, although, as previously stated (see Section 8.3.2), this approach has many potential problems.

Rather than removing variables or units with missing data, we can impute missing values. We handle the problem of missing data in the ABC ACSS using both single and multiple imputation. We initially use the naive unconditional mean imputation approach to impute one value for each missing item: missing values in each satisfaction variable are filled in by the median of the recorded values of that variable. The median, rather than the mean, is used because each satisfaction item is an ordinal categorical variable. Unfortunately, this strategy can severely distort the distribution for variables, leading to complications with summary measures including, most notably, underestimation of variances. Moreover, this method will distort associations between variables.

As an alternative to unconditional median imputation, we use the R environment and functions from the `mice` library to create $m = 5$ multiply-imputed data sets for each of three multiple-imputation methods. The first imputes the missing values in each variable using a simple random sample from the observed values of that variable. This method is useful if the data are assumed to be MCAR, but it does not account for the associations between variables. The second method fills in missing values using Bayesian polytomous logistic regression for nominal and ordinal categorical variables, and Bayesian linear regression for quantitative variables. The third method is like the second, but ordinal categorical variables (such as the satisfaction items) are modeled using Bayesian linear regression, and the imputed values are then rounded to the nearest level observed in the data.

In our simple application, we focus on two estimands: the proportion of satisfied or highly satisfied (satisfaction level ≥ 4) customers for each satisfaction item, and the association between the overall satisfaction variables (overall satisfaction with ABC, willingness to recommend ABC to other companies, and repurchasing intentions) and satisfaction with sales support, technical support, and software solutions. Association is measured using

Table 8.1 Univariate analyses of ABC ACSS data using complete-case analysis (CCA), available-case analysis (ACA), single imputation (SI) and multiple imputation (MI): proportion of highly or very highly satisfied customers (standard errors in parentheses)

	Missing-data					MI‡	
Variable	Proportion	CCA	ACA	SI†	(*a*)	(*b*)	(*c*)
Overall satisfaction with ABC							
Overall satisfaction	0.015	0.507	0.595	0.602	0.592	0.590	0.592
		(0.061)	(0.030)	(0.030)	(0.031)	(0.030)	(0.031)
Willingness to recommend ABC	0.019	0.537	0.632	0.639	0.632	0.629	0.630
		(0.061)	(0.030)	(0.029)	(0.030)	(0.030)	(0.030)
Repurchasing intentions	0.015	0.552	0.649	0.654	0.647	0.644	0.647
		(0.061)	(0.029)	(0.029)	(0.029)	(0.029)	(0.030)
Overall satisfaction with …							
Equipment	0.034	0.657	0.623	0.635	0.623	0.616	0.617
		(0.058)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)
Sales support	0.068	0.493	0.472	0.440	0.464	0.471	0.471
		(0.061)	(0.032)	(0.030)	(0.033)	(0.031)	(0.033)
Technical support	0.008	0.716	0.686	0.688	0.683	0.685	0.683
		(0.055)	(0.029)	(0.028)	(0.029)	(0.029)	(0.029)
Training	0.338	0.657	0.733	0.823	0.737	0.709	0.714
		(0.058)	(0.033)	(0.023)	(0.034)	(0.035)	(0.035)
Supplies	0.053	0.388	0.472	0.447	0.471	0.467	0.471
		(0.060)	(0.031)	(0.030)	(0.031)	(0.032)	(0.032)
Software solutions	0.312	0.388	0.393	0.271	0.401	0.375	0.371
		(0.060)	(0.036)	(0.027)	(0.032)	(0.036)	(0.030)
Customer website	0.214	0.433	0.455	0.357	0.450	0.432	0.459
		(0.061)	(0.034)	(0.029)	(0.033)	(0.036)	(0.038)
Purchasing support	0.086	0.537	0.519	0.560	0.531	0.517	0.511
		(0.061)	(0.032)	(0.030)	(0.032)	(0.032)	(0.032)
Contracts and pricing	0.049	0.224	0.332	0.316	0.334	0.325	0.328
		(0.051)	(0.030)	(0.029)	(0.031)	(0.029)	(0.029)

†Unconditional median imputation was used.

‡Three imputation methods were used: (a) imputations are randomly drawn from the observed values; (b) a separate univariate regression model is specified for each variable, taking into account the measurement scale of that variable (Bayesian polytomous logistic regression for categorical variables and Bayesian linear regression for quantitative variables); (c) as with (b), but ordinal categorical variables are modeled using Bayesian linear regression, and the imputed values are rounded to the nearest level observed in the data.

the Goodman–Kruskal gamma statistic. Table 8.1 (last six columns) and Table 8.2 present estimates and their standard errors for these univariate and bivariate statistics, using the alternative missing-data methods mentioned above.

As can be seen in these tables, the standard error estimates for the complete-case estimators are twice as large as those of the available-case analysis and the single and multiple imputation-based analyses, showing the large loss of precision due to discarding incomplete cases. Moreover, the differences between the complete-case and available-case estimates

Table 8.2 Bivariate analyses of ABC ACSS data using complete-case analysis (CCA), available-case analysis (ACA), single imputation (SI) and multiple imputation (MI): the Goodmann–Kruskal gamma statistic (standard errors in parentheses)

Target variable: overall satisfaction with ABC						
					MI‡	
Variable	CCA	ACA	SI†	(*a*)	(*b*)	(*c*)
Willingness to recommend ABC	0.889	0.890	0.893	0.883	0.894	0.889
	(0.047)	(0.025)	(0.025)	(0.026)	(0.024)	(0.025)
Repurchasing intentions	0.676	0.764	0.766	0.749	0.766	0.763
	(0.095)	(0.039)	(0.039)	(0.045)	(0.039)	(0.039)
Overall satisfaction with . . .						
Sales support	0.541	0.418	0.409	0.381	0.390	0.382
	(0.124)	(0.068)	(0.068)	(0.069)	(0.069)	(0.068)
Technical support	0.595	0.646	0.636	0.631	0.650	0.641
	(0.115)	(0.051)	(0.052)	(0.057)	(0.051)	(0.053)
Software Solutions	0.372	0.487	0.460	0.353	0.401	0.413
	(0.137)	(0.078)	(0.073)	(0.079)	(0.081)	(0.077)
Target variable: willingness to recommend ABC to other companies						
Overall satisfaction	0.889	0.890	0.893	0.883	0.894	0.889
	(0.047)	(0.025)	(0.025)	(0.026)	(0.024)	(0.025)
Repurchasing intentions	0.767	0.889	0.890	0.876	0.889	0.888
	(0.081)	(0.028)	(0.028)	(0.030)	(0.028)	(0.028)
Overall satisfaction with . . .						
Sales support	0.583	0.457	0.438	0.410	0.434	0.434
	(0.105)	(0.061)	(0.061)	(0.063)	(0.063)	(0.061)
Technical support	0.409	0.481	0.474	0.473	0.483	0.473
	(0.126)	(0.058)	(0.057)	(0.059)	(0.057)	(0.060)
Software solutions	0.389	0.496	0.480	0.349	0.433	0.478
	(0.127)	(0.076)	(0.072)	(0.084)	(0.067)	(0.063)
Target variable: repurchasing intentions						
Overall satisfaction	0.676	0.764	0.766	0.749	0.766	0.763
	(0.095)	(0.039)	(0.039)	(0.045)	(0.039)	(0.039)
Willingness to recommend ABC	0.767	0.889	0.890	0.876	0.889	0.888
	(0.081)	(0.028)	(0.028)	(0.030)	(0.028)	(0.028)
Overall satisfaction with . . .						
Sales support	0.675	0.477	0.463	0.430	0.459	0.457
	(0.082)	(0.059)	(0.059)	(0.063)	(0.061)	(0.059)
Technical support	0.222	0.381	0.378	0.380	0.390	0.376
	(0.140)	(0.065)	(0.065)	(0.066)	(0.065)	(0.067)
Software solutions	0.336	0.416	0.412	0.301	0.353	0.398
	(0.134)	(0.082)	(0.077)	(0.079)	(0.071)	(0.070)

†Unconditional median imputation was used.

‡Three imputation methods were used: (a) imputations are randomly drawn from the observed values; (b) a separate univariate regression model is specified for each variable, taking into account the measurement scale of that variable (Bayesian polytomous logistic regression for categorical variables and Bayesian linear regression for quantitative variables); (c) as with (b), but ordinal categorical variables are modeled using Bayesian linear regression, and the imputed values are rounded to the nearest level observed in the data.

suggest that the MCAR assumption is not plausible for the 2010 ACSS data, and so complete-case estimates may be biased. As we might expect, multiple-imputation methods lead to standard error estimates slightly larger than those of single imputation, because they incorporate all uncertainty due to predicting missing data.

Consider now the univariate statistics in Table 8.1. The differences among the point estimates provided by the alternative missing-data methods are not striking irrespective of the proportion of missing values in each item. However, complete-case and available-case analyses, as well as unconditional median imputation, cannot generally be recommended, given their theoretical pitfalls.

Larger differences between inferences provided by the alternative missing-data methods can be observed for the bivariate analyses. Specifically, the two multiple-imputation methods, which incorporate information on the relationships between variables (last two columns of Table 8.2), lead to quite different results than the marginal multiple-imputation method (column named MI (a)) and the other more naive approaches. However, these results do not appear to be sensitive to the models used for imputation represented in the last two columns.

As a closing note, we offer two reminders. First, as previously noted, missing data in the ABC ACSS are not monotone, and therefore the MICE strategy we applied might theoretically use incompatible fully conditional distributions for imputing missing values. Alternative imputation methods, such as IMB (e.g., Baccini *et al.*, 2010), could be used and the results compared with those from MICE. Second, although not a focus of this review, sensitivity analyses are an important component of modeling when relatively many data are missing, and they should be routinely conducted (see also Chapter 9 of this book on dealing with problems of outlier detection). Such additional analyses require effort, but allow insight into the impact of missing data assumptions (e.g., Baccini *et al.*, 2009).

8.8 Summary

Missing data are a prevalent problem in many social and economic studies, including customer satisfaction surveys. Here, we have reviewed some important general concepts regarding missing-data patterns and mechanisms that lead to missing data, and then discussed some common, but naive, techniques for dealing with missing data, as well as less naive and more principled methods. Simple approaches, such as complete-case analysis and available-case analysis, provide inefficient, though valid, results when missing data are MCAR, but generally biased results when missing data are MAR. Other frequently used methods for handling missing data, such as unconditional mean imputation, provide generally biased results even under the MCAR assumption.

Multiple imputation is a principled method and a useful tool for handling missing data because of its flexibility allowing the imputation and subsequent analysis to be conducted separately. This key feature of MI is especially attractive in the context of public-use data sets to be shared by many users. However, the validity of the analysis can depend on the cability of the imputation model to capture the missingness mechanism correctly. Also, the separation of imputation model and analysis model raises the issue of compatibility between these two models. If an appropriate imputation model is employed, MI generally leads to statistically valid inferences. MI is also becoming popular because of the availability of easy-to-use statistical software. Such software helps users to apply MI in a broad range of missing-data settings.

MI is not the only principled method for handling missing values, nor is it necessarily the best for a specific problem. In some cases, good estimates can be obtained through a weighted estimation procedure. In fully parametric models, ML or Bayesian inferences can often be conducted directly from the incomplete data by specialized numerical methods, such as the EM or DA algorithms, or their extensions. MI has the advantages of flexibility over direct analyses, by allowing the imputer's and analyst's models to differ.

A crucial issue arising in the analysis of incomplete data concerns the uncertainty about the reasons for nonresponse. Therefore it is useful to conduct sensitivity analyses under different modeling assumptions. Multiple imputation allows the straightforward study of the sensitivity of inferences to various missing-data mechanisms simply by creating repeated randomly drawn imputations under more than one model and using complete-data methods repeatedly (e.g., Rässler, 2002; Rubin, 1977, 1986; Baccini *et al.*, 2009, 2010).

Many of the approaches discussed here may be applied under either an ignorable or nonignorable model for the missing-data mechanism. The observed data can never provide any direct evidence against ignorability, and procedures based on ignorable missing-data models typically lead to at least partial corrections for the bias due to nonresponse.

Acknowledgements

This chapter borrows in some places from previous reviews written by other combinations of co-authors. These reviews are presented in publications by Cook and Rubin (2005), Rässler *et al*. (2007) and Rässler *et al*. (2008). We thank the authors of these papers for their generosity.

References

Baccini, M., Cook, S., Frangakis, C.E., Li, F., Mealli, F., Rubin, D.B. and Zell, E.R. (2009) Evaluating multiple imputation procedures using simulations in a Bayesian prospective. Paper presented at the 2009 Joint Statistical Meetings, Washington, DC.

Baccini, M., Cook, S., Frangakis, C.E., Li, F., Mealli, F., Rubin, D.B. and Zell, E.R. (2010) Multiple imputation in the anthrax vaccine research program. *Chance*, 23, 16–23.

Barnard, J. and Rubin, D.B. (1999) Small-sample degrees of freedom with multiple imputation, *Biometrika*, 86, 948–955.

Bethlehem, J.G. (2002) Weighting nonresponse adjustments based on auxiliary information. In R.M. Groves, D.A. Dillman, J.L. Eltinge and R.L.A. Little (eds), *Survey Nonresponse*, pp. 275–287. New York: John Wiley & Sons, Inc.

Cook, S. and Rubin, D.B. (2005) Multiple imputation in designing medical device trials. In K.M. Becker and J.J. Whyte (eds), *Clinical Evaluation of Medical Devices*, pp. 241–251. Washington, DC: Humana Press.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–22.

Efron, B. (1994) Missing data, imputation, and the bootstrap, *Journal of the American Statistical Association*, 89, 463–475.

Fay, R.E. (1992) When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 227–232.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. and Carlin, J.B. (2002) Poststratification and weighting adjustment. In R.M. Groves, D.A. Dillman, J.L. Eltinge and R.L.A. Little (eds), *Survey Nonresponse*, pp. 289–302. New York: John Wiley & Sons, Inc.

Gelman, A., Carlin J.B., Stern, H.S. and Rubin, D.B. (2003) *Bayesian Data Analysis*, 2nd edn. Boca Raton, FL: Chapman & Hall.

Gelman, A., Hill, J., Yajima, M., Su, Y. and Pittau, M. (2011) mi: Missing data imputation and model checking. Package for the R statistical software. http://lib.stat.cmu.edu/R/CRAN/.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J.E. (1996) *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (1986) Selection modeling versus mixture modeling with noningnorable nonresponse. In H. Wainer (ed.), *Drawing Inferences from Self-Selected Samples*, pp. 115–142. New York, Springer.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (1993) Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*, 88, 984–993.

Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (eds) (2002) *Survey Nonresponse*. New York: John Wiley & Sons, Inc.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

Heckman, J.J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475–492.

Horton, N.J. and Kleinman, K.P. (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistician*, 61(1), 79–90.

Horton, N.J. and Lipsitz, S.R. (2001) Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician*, 55, 244–254.

Kennickell, A.B. (1991) Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 1–10.

Kim, J.K., Brick, J.M., Fuller, W.A. and Kalton, G. (2006) On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of The Royal Statistical Society, Series B*, 68(3), 509–521.

Lee, H., Rancourt, E. and Särndal, C.E. (2002) Variance estimation for survey data under single imputation. In R.M. Groves, D.A. Dillman, J.L. Eltinge and R.L.A. Little (eds), *Survey Nonresponse*, pp. 315–328. New York: John Wiley & Sons, Inc.

Li, K.H., Meng, X.-L., Raghunathan, T.E. and Rubin, D.B. (1991a) Significance levels from repeated *p*-values with multiply-imputed data. *Statistica Sinica*, 1, 65–92.

Li, K.H., Raghunathan, T.E. and Rubin, D.B. (1991b) Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86, 1065–1073.

Little, R.J.A. (1985) A note about models for selectivity bias. *Econometrica*, 53, 1469–1474.

Little, R.J.A. (1993) Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88, 125–134.

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

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. Hoboken, NJ: John Wiley & Sons, Inc.

Little, R.J.A. and Schenker, N. (1995) Missing data. In G. Arminger, C.C. Clogg and M.E. Sobel (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, pp. 39–75. New York: Plenum Press.

Liu, C. and Rubin, D.B. (1994) The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81, 533–648.

Liu, C., Rubin, D.B. and Wu, Y.N. (1998) Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, 85, 755–770.

Madaw, W.G. and Olkin, I. (1983) *Incomplete Data in Sample Surveys. Volume 3: Proceedings of the Symposium*. New York: Academic Press.

Madaw, W.G., Nisselson, H. and Olkin, I. (1983a) *Incomplete Data in Sample Surveys. Volume 1: Report and Case Studies*. New York: Academic Press.

Madaw, W.G., Olkin, I. and Rubin, D.B. (1983b) *Incomplete Data in Sample Surveys. Volume 2: Theory and Bibliographies and Case Studies*. New York: Academic Press.

McLachlan, G.J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. New York: John Wiley & Sons, Inc.

Meng, X.-L. (1994) Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 10, 538–573.

Meng, X.-L. and Rubin, D.B. (1992) Performing likelihood ratio tests with multiply imputed data sets. *Biometrika*, 79, 103–111.

Meng, X.-L. and Rubin, D.B. (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267–278.

Meng, X.-L. and van Dyk, D. (1997) The EM algorithm: An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B*, 59, 511–567.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1091.

Metropolis, N. and Ulam, S. (1949) The Monte Carlo method. *Journal of the American Statistical Association*, 49, 335–341.

Münnich, R. and Rässler, S. (2005) PRIMA: A new multiple imputation procedure for binary variables. *Journal of Official Statistics*, 21, 325–341.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85–95.

Rässler, S. (2002) *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Lecture Notes in Statistics 168. New York: Springer.

Rässler, S., Rubin, D.B. and Schenker, N. (2007) Incomplete data: Diagnosis, imputations, and estimation. In E. de Leeuw, J. Hox and D. Dillman (eds), *The International Handbook of Survey Research Methodology*. Thousand Oaks, CA: Sage.

Rässler, S., Rubin, D.B. and Zell, E.R. (2008) Incomplete data in epidemiology and medical statistics. In C.R. Rao, J.P. Miller and D.C. Rao (eds), *Handbook of Statistics*, 27, pp. 569-601. Amsterdam: Elsevier.

Royston, P. (2004) Multiple imputation of missing values. *Stata Journal*, 4(3), 227–241.

Royston, P. (2005) Multiple imputation of missing values: update. *Stata Journal*, 5(2), 188–201.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63(3), 581–592.

Rubin, D.B. (1977) Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 538–543.

Rubin, D.B. (1978a) Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20–34.

Rubin, D.B. (1978b) A note on Bayesian, likelihood, and sampling distribution inferences. *Journal of Educational Statistics*, 3, 189–201.

Rubin, D.B. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4, 87–95.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1994) Comment on ‘Missing data, imputation, and the bootstrap’ by B. Efron, *Journal of the American Statistical Association*, 89, 475–478.

Rubin, D.B. (1996) Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473–489.

Rubin, D.B. (2003) Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* 57(1), 3–18.

Rubin, D.B. (2004a) *Multiple Imputation for Nonresponse in Surveys*, 2nd edn. Hoboken, NJ: John Wiley & Sons, Inc.

Rubin, D.B. (2004b) The design of a general and flexible system for handling nonresponse in sample surveys. *American Statistician*, 58, 298–302.

Rubin, D.B. and Schenker, N. (1986) Multiple imputation for interval estimation from simple random sample with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366–374.

Rubin, D.B. and Schenker, N. (1991) Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine*, 10, 585–598.

Rubin, D.B., Stern, H. and Vehovar, V. (1995) Handling ‘don’t know’ survey responses: The case of Slovenian plebiscite. *Journal of the American Statistical Association*, 90, 822–828.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

Schafer, J.L. and Schenker, N. (2000) Inference with imputed conditional means, *Journal of the American Statistical Association*, 95, 144–154.

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. and Rubin, D.B. (1996) The NHANES III Multiple Imputation Project. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Shao, J. (2002) Replication methods for variance estimation in complex surveys with imputed data. In R.M. Groves, D.A. Dillman, J.L. Eltinge and R.L.A. Little (eds), *Survey Nonresponse*, pp. 303–314. New York: John Wiley & Sons, Inc.

Shih, W.J. (1992) On informative and random dropouts in longitudinal studies. *Biometrics*, 48, 970–972.

Su, Y.-S., Gelman, A., Hill, J., Yajima, M. (forthcoming) Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*. http://www.stat.columbia.edu/gelman/research/published/mipaper.rev04.pdf.

Tang, L., Song, J., Belin, T.R. and Unuetzer, J. (2005) A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine*, 24, 2111–2128.

Tanner, M.A. and Wong, W.H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528–550.

Van Buuren, S. and Groothuis-Oudshoorn K. (forthcoming) MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software.* http://lib.stat.cmu.edu/R/CRAN/web/packages/mice/index.htm.

Van Buuren, S. and Oudshoorn, C.G.M. (2000) *Multivariate Imputation by Chained Equations: MICE v1.0 User's Manual*, Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid.

Van Buuren, S. and Oudshoorn, K. (1999) *Flexible Multivariate Imputation by MICE*, TNO/VGZ/PG 99.054. Leiden: TNO Preventie en Gezondheid.

Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. and Rubin, D.B. (2006) Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76, 1049–1064.