Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables

Nicholas J. HORTON and Stuart R. LIPSITZ

Missing data frequently complicates data analysis for scientific investigations. The development of statistical methods to address missing data has been an active area of research in recent decades. Multiple imputation, originally proposed by Rubin in a public use dataset setting, is a general purpose method for analyzing datasets with missing data that is broadly applicable to a variety of missing data settings. We review multiple imputation as an analytic strategy for missing data. We describe and evaluate a number of software packages that implement this procedure, and contrast the interface, features, and results. We compare the packages, and detail shortcomings and useful features. The comparisons are illustrated using examples from an artificial dataset and a study of child psychopathology. We suggest additional features as well as discuss limitations and cautions to consider when using multiple imputation as an analytic strategy for incomplete data settings.

KEY WORDS: Generalized linear models; Incomplete data; Markov Chain Monte Carlo; Missing outcomes; Missing predictors.

1. INTRODUCTION

Missing data is a commonly occurring complication in many scientific investigations. Determining the appropriate analytic approach in the presence of incomplete observations is a major question for data analysts. The development of statistical methods to address missing data has been an active area of research in recent decades (Rubin 1976; Little and Rubin 1987; Laird 1988; Ibrahim 1990; Little 1992; Robins, Rotnitzky, and Zhao 1994, 1995; Horton and Laird 1999). There are three types of concerns that typically arise with missing data: (1) loss of efficiency; (2) complication in data handling and analysis; and (3) bias due to differences between the observed and unobserved data (Barnard and Meng 1999). One approach to incomplete data problems that addresses these concerns is multiple imputation, which was proposed 20 years ago by Rubin (1977) and described in detail by Rubin (1987) and Schafer (1997). A concise and readable primer can be found in Schafer (1999), while Rubin (1996) provided an extensive bibliography.

This article reviews multiple imputation, describes assumptions that it requires, and reviews software packages that implement this procedure. We apply the methods and compare the results using two examples—a child psychopathology dataset with missing outcomes and an artificial dataset with missing covariates. We conclude with some discussion of the strengths and weaknesses of these implementations as well as advantages and limitations of imputation.

Nicholas J. Horton is Assistant Professor, Department of Epidemiology and Biostatistics, Boston University School of Public Health, and Department of Medicine, Boston University School of Medicine, 715 Albany Street, T3E, Boston, MA 02118 (E-mail: horton@bu.edu). Stuart R. Lipsitz is Professor, Department of Biometry and Epidemiology, Medical University of South Carolina, Charleston, SC 29425. The authors are grateful for the support provided by NIMH grant R01-MH546932. We also thank Gwendolyn Zahner for use of the child psychopathology dataset, which was conducted under contract to the Connecticut Department of Children and Youth Services while Dr. Zahner was on the faculty of the Yale University Child Study Center.

2. MULTIPLE IMPUTATION

Rubin (1996) described multiple imputation as a three-step process. First, sets of plausible values for missing observations are created that reflect uncertainty about the nonresponse model. Each of these sets of plausible values can be used to “fill-in” the missing values and create a “completed” dataset. Second, each of these datasets can be analyzed using complete-data methods. Finally, the results are combined, which allows the uncertainty regarding the imputation to be taken into account.

The method of multiple imputation was first proposed in a public-use survey data setting. Consider an intensive survey of a small geographical area, where some observations were missing or incomplete. Other researchers, whom we denote as *users* of the survey, may have interest in these data. Care needs to be taken by the *creators* of the survey to ensure that respondents are not uniquely identified based on characteristics included in the public dataset. Note that disclosure of the identity of the respondents may occur if too much information (such as area, gender, ethnicity, occupation, age, and so on) is provided regarding the sample (Fienberg 1994; Fienberg and Willenborg 1998; Zaslavsky and Horton 1998).

Multiple imputation remains ideally suited to this setting, since the *creators* can use auxiliary confidential and detailed information that would be inappropriate to include in the public dataset (Rubin 1996). This comprehensive set of information about the study can be used to “fill-in” or impute sets of values for incomplete observations. Given the complete datasets, *users* may utilize existing software to analyse each of the datasets. Finally, given the results for each analysis, an overall summary is straightforward to calculate.

Although multiple imputation was initially proposed by Rubin for public use data, its use has broadened to general purpose missing data settings. Here we consider general regression models with outcomes (denoted by $\mathbf{Y}$, which may be scalar or vector valued) and a vector of predictors (denoted by $\mathbf{X}$). For a given subject, these quantities are either observed or missing. We denote $\mathbf{Y}^{\text{obs}}$ as the observed component of the outcome and $\mathbf{X}^{\text{obs}}$ as the observed components of the predictors. Similarly, we denote $\mathbf{Y}^{\text{mis}}$ and $\mathbf{X}^{\text{mis}}$ as the unobserved components of the outcome and predictors, respectively. We will also refer to $\mathbf{Z}^{\text{mis}} = (\mathbf{Y}^{\text{mis}}, \mathbf{X}^{\text{mis}})$ and $\mathbf{Z}^{\text{obs}} = (\mathbf{Y}^{\text{obs}}, \mathbf{X}^{\text{obs}})$.

Primary interest revolves around the regression parameters $\boldsymbol{\beta}$ governing the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$: $f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta})$, and we are concerned with efficiency issues, complications in analysis, and bias. Before describing multiple imputation in detail, we will review classifications for the probability distribution generating the missing data, using the nomenclature of Little and Rubin (1987).

2.1 Missing Data Classifications

Let $\mathbf{Z}$ be partially observed, where $\mathbf{R}$ is a set of response indicators (i.e., $R_j = 1$ if the jth element of $\mathbf{Z}$ is observed, and equals 0 otherwise), governed by parameters $\boldsymbol{\phi}$. The missing completely at random (MCAR) assumption is defined as

$$P(\mathbf{R}|\mathbf{Z}) = P(\mathbf{R}|\mathbf{Z}^{\text{obs}}, \mathbf{Z}^{\text{mis}}) = P(\mathbf{R}|\boldsymbol{\phi}),$$

where in addition $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ are presumed distinct. Heuristically, this assumption states that missingness is not related to any factor, known or unknown, in the study.

It may be more plausible to posit that missingness is missing at random (MAR), which assumes that

$$P(\mathbf{R}|\mathbf{Z}) = P(\mathbf{R}|\mathbf{Z}^{\text{obs}}, \boldsymbol{\phi}).$$

Heuristically, this states that the missingness depends only on observed quantities, which may include outcomes, and a rich set of predictors. It is possible to statistically test the MCAR assumption, against the alternate hypothesis that missingness is MAR (Diggle, Liang, and Zeger 1994; Little 1988).

Finally, if the missingness law $P(\mathbf{R}|\mathbf{Z})$ cannot be simplified (i.e., it depends on unobserved quantities), the process is termed *nonignorable*. In a nonignorable nonresponse setting, the correct specification of the missingness law must be given to yield consistent estimates of the regression parameters. Without additional information, it is impossible to test the MAR assumption against a nonignorable alternative (Little and Rubin 1987).

	Monotone			*Nonmonotone*		
Obs	Z_1	Z_2	Z_3	Z_1	Z_2	Z_3
1	O	O	O	O	M	O
2	O	M	M	O	M	M
3	O	M	M	O	M	M
4	O	M	M	O	M	M
5	O	O	M	O	O	M
6	O	O	M	O	O	M
7	O	O	M	O	O	M
8	O	O	O	O	O	O

Figure 1. Monotone and Nonmonotone Patterns of Missingness (O = observed, M = missing).

Another important distinction regarding the missing data refers to the pattern of missing data. If the data matrix can be rearranged in such a way that there is a hierarchy of missingness, such that observing a particular variable Z_b for a subject implies that Z_a is observed, for $a < b$, then the missingness is said to be *monotone*. Simpler imputation methods can be used if the pattern is monotone, though a monotone pattern is uncommon in most complex investigations. It may be possible, however, to create a monotone missingness pattern that separates out a small number of observations that are non-monotone. In the nonmonotone example in Figure 1, all but the first observation can be rearranged into a monotone pattern (i.e., 83% of the dataset). This type of hybrid pattern is exploited by a number of computer packages, since this allows a much simpler model to be estimated.

It is important to note that in many realistic settings, datasets may have missing outcomes as well as missing predictors (Little 1992). Such patterns tend to complicate analysis, particularly for regression models with many predictors.

2.2 Steps for Multiple Imputation

More formally, the three steps for multiple imputation consist of:

Imputation: Generate a set of $m > 1$ plausible values for $\mathbf{Z}^{\text{mis}} = (\mathbf{Y}^{\text{mis}}, \mathbf{X}^{\text{mis}})$.
Analysis: Analyze the m datasets using complete-case methods.
Combination: Combine the results from the m analyses.

Imputation step—The imputation step is perhaps most critical, since it relies upon assumptions regarding the missingness law that generated the observed sample. The goal of the imputation is to account for the relationships between the unobserved and observed variables, while taking into account the uncertainty of the imputation. The MAR assumption (which is generally assumed for many missing data methods, and as previously noted is untestable without additional information) is key to the validity of multiple imputation. Use of this assumption allows the analyst to generate imputations $(\mathbf{Z}^{\{1\}}, \mathbf{Z}^{\{2\}}, \ldots \mathbf{Z}^{\{m\}})$ from the distribution $f(\mathbf{Z}^{\text{mis}}|\mathbf{Z}^{\text{obs}})$, since after conditioning on $\mathbf{Z}^{\text{obs}}$, missingness is assumed to be due to chance.

There are a variety of imputation models that have been used. When missingness is monotone, simple methods have been proposed, including (for continuous variables) propensity methods (Rosenbaum and Rubin 1983), predictive mean matching (Little 1988), and (for discrete variables) discriminant analysis or logistic regression. For more complicated missingness, Markov Chain Monte Carlo (MCMC) approaches have been suggested. Both the predictive mean matching and MCMC approaches require assumptions of multivariate normality, but there is some evidence (see, e.g., Schafer 1997) that the inferences tend to be robust to minor departures from this assumption.

As an example of an imputation model, we will describe the predictive mean matching approach, where a linear regression is postulated for the distribution of a partially observed variable, conditional on other factors. For the predictive mean matching approach with a variable Z_j with missing values, a model is fit using complete observations for $Z_1, \ldots, Z_{j-1}$:

$$E[Z_j|\boldsymbol{\phi}] = \phi_0 + \phi_1 Z_1 + \phi_2 Z_2 + \cdots + \phi_{j-1} Z_{j-1}. \quad (1)$$

Next, new parameters ϕ^* are drawn from the distribution of the parameters (since these values were estimated and not known with certainty). Finally, for the ith imputation, the missing values are replaced by:

$$Z_j^{(i)} = \phi_0^* + \phi_1^* Z_1 + \phi_2^* Z_2 + \cdots + \phi_{j-1}^* Z_{j-1} + \sigma^* \epsilon,$$

where σ^* is the estimate of variance from the model and ϵ is a simulated normal random variate. We refer to this as the *regression* method. A variant of this approach imputes the observed value of Z_j that is closest to $\hat{Z}_j$ in the dataset; this ensures that imputed values are plausible, and may be more appropriate if the normality assumption is violated. We refer to this as the *predictive mean matching method*. The predictive mean model assumes a linear regression model and a monotone structure, otherwise there will be missing predictors in model (1).

The propensity score method uses a different model for imputation. Here, values are imputed from observations that are equally likely to be missing, by fitting a logistic regression model for the missingness indicators. Allison (2000) noted that this method can yield serious bias for imputation of missing covariates under some settings.

For discrete incomplete variables, discriminant analysis (or logistic regression for dichotomous variables) can be used to impute values based on the estimated probability that a missing value takes on a certain value $P(Z_j^{\text{mis}} = k|\mathbf{Z}^{\text{obs}})$. Using Bayes's theorem, this can be calculated from estimates of the joint distribution of $\mathbf{Z}$.

Finally, the MCMC method constructs a Markov chain to simulate draws from the posterior distribution of $f(\mathbf{Z}^{\text{mis}}|\mathbf{Z}^{\text{obs}})$. This can be implemented using the IP algorithm (Schafer 1997), where at the tth iteration the steps can be defined as:

Imputation-step: Draw $\mathbf{Z}^{\text{mis},(t+1)}$ from $f(\mathbf{Z}|\mathbf{Z}^{\text{obs}}, \boldsymbol{\phi}^{(t)})$.

Parameter-step: Draw $\boldsymbol{\phi}^{(t+1)}$ from $f(\boldsymbol{\phi}|\mathbf{Z}^{obs}, \mathbf{Z}^{\text{mis},(t+1)})$.

The Markov chain

$$\left(\left\{\mathbf{Z}^{(1)}, \boldsymbol{\phi}^{(1)}\right\}, \left\{\mathbf{Z}^{(2)}, \boldsymbol{\phi}^{(2)}\right\}, \ldots, \left\{\mathbf{Z}^{(t+1)}, \boldsymbol{\phi}^{(t+1)}\right\}, \ldots\right)$$

can be shown to converge to the posterior distribution of interest. This method has the advantage that it can handle arbitrary patterns of missing data. Schafer (1997) provided a complete exposition of the method in the imputation setting, while Gilks, Richardson, and Spiegelhalter (1996) described the background as well as other applications. As a computational tool MCMC has some downsides; it requires an assumption of multivariate normality, it is complicated and computationally expensive. Convergence is difficult to determine, and remains more of an art form than a science. However, MCMC is available in SAS, S-Plus, and MICE, and thus is becoming more mainstream.

Some practical suggestions for what variables to include in the imputation model were given by van Buuren, Boshuizen, and Knook (1999). They recommended that this set of variables includes those in the complete data model, factors known to be associated with missingness, and factors that explain a considerable amount of variance for the target variables.

"Complete data" analysis step—The next step in the imputation process is to carry out the analysis of interest for each of the m imputed complete-observation datasets, storing the parameter vector and standard error estimates.

Combination step—Finally, the results are combined using results from Rubin (1987), to calculate estimates of the within imputation and between imputation variability. These statistics account for the variability of the imputations and assuming that the imputation model is correct, provide consistent estimates of the parameters and their standard errors. There has been an extensive literature regarding the asymptotic behavior of multiple imputation methods (Barnard and Rubin 1999; Meng and Rubin 1992; Robins and Wang 2000; Rubin 1996; Wang and Robins 1998) these issues are not further considered here.

Notes about imputation—It should be noted that one advantage of multiple imputation as an analytic approach is that it allows the analyst to incorporate additional information into the imputation model. This auxiliary (or extraneous) information may not be of interest in the regression model, but may make the MAR assumption increasingly plausible (Liu, Taylor, and Belin 2000; Rubin 1996); such information is straightforward to incorporate into the imputation model.

A useful quantity in interpreting results from multiple imputation is an estimate of the fraction of missing information (Rubin 1987). This quantity, typically denoted by $\hat{\gamma}$, denotes how the missing data influence the uncertainty of estimates of $\boldsymbol{\beta}$ (Schafer 1997). It has been noted that even with a large fraction of missing information, a relatively small number of imputations provides estimates of standard errors that are almost fully efficient (Schafer 1997). Schafer (1999) suggested that no more than 10 imputations are usually required, though this should be investigated more closely if the fraction of missing information is large. In any case, the appropriate number of imputations can be informally determined by carrying out replicate sets of m imputations and determining whether the estimates are stable between sets.

Before the advent of general purpose packages that supported multiple imputation, the process of generating imputed datasets, managing the results from each of the m datasets, and combining the results required specialized programming or use of macros that were difficult to use. The packages reviewed in this paper, though still more complicated than complete case methods, greatly facilitate the process of using multiple imputation.

3. SOFTWARE PACKAGES

SOLAS version 3.0
Statistical Solutions (North American office)
Stonehill Corporate Center, Suite 104,
999 Broadway, Saugus, MA 01906, USA.
Tel (781) 231-7680
http://www.statsol.ie/solas/solas.htm, sales@statsol.ie

SOLAS is designed specifically for the analysis of datasets with missing observations. It offers a variety of multiple imputation techniques in a single, easy-to-use package with a well-designed graphical user interface.

SOLAS supports predictive mean model (using the closest observed value to the predicted value) and propensity score models for missing continuous variables, and discriminant models for missing binary and categorical variables. Once the multiple datasets are created, the package allows the calculation of descriptive statistics, t tests and ANOVA, frequency tables, and linear regression.

The system automatically provides summary measures by combining the results from the multiple analyses. These reports can be saved as rich text format (rtf) files. It has extensive capabilities to read and write database formats (1-2-3, dBase, FoxPro, Paradox), spreadsheets (Excel), and statistical packages (Gauss, Minitab, SAS, S-Plus, SPSS, Stata, Statistica, and Systat). Imputed datasets can be exported to other statistical packages, though this is a somewhat cumbersome process, since the combination of results from the multiple analyses then needs to be done manually.

The new script language facility is particularly useful in documenting the steps of a multiple imputation run, and for conducting simulations. It can be set up to automatically record the settings from a menu-based multiple imputation session, and store this configuration in a file for later revision and reuse.

SOLAS includes the ability to view missing data patterns, review the quantity and positioning of missing values, and classify them into categories of monotone or nonmonotone missingness. Because it does not consolidate observations with the same pattern of missing data, however, this feature is of limited utility in large datasets.

A nice feature of SOLAS is the fine-grained and intuitive control of the details of the imputation model. As an example, incorporating auxiliary information (variables in the imputation model but not in the regression model of interest) is straightforward.

There are a number of limitations to SOLAS' implementation. It is primarily a package for multiple imputation in linear regression models, and has limited data manipulation ability. While there are extensive options for linear regression, it lacks the completeness of general-purpose packages (e.g., specification of interactions is cumbersome). Nonlinear regression methods, such as logistic or survival models, are not supported. Nonmonotone missingness is handled in an ad-hoc fashion. While this may be acceptable in many applications, it is not always appropriate (support for MCMC methods, not available in SOLAS, are particularly attractive in this setting). By default, SOLAS does not display any estimates of the fraction of missing information (this can be calculated separately from the regression model), nor standard error estimates for the intercept. The default behavior uses a fixed seed for the random number generator seed. While this can be set to 0 (or left blank) to use clock time as seed (except within the script language), the default seed will always yield the same imputation results. Using the clock time as a pseudo-random seed would seem a more reasonable default.

Installation was straightforward, and the interface was clear and intuitive. The documentation (which consisted of three manuals, with a total of 487 pages) was well organized (it was the only documentation with an index), and easy to read. The extensive examples with numerous screen-shots were useful in introducing the user to SOLAS.

A single user license for SOLAS 3.0 costs $1,295 ($995 for academic customers), while an upgrade from previous versions costs $495.

SAS 8.2 (beta) SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513-2414
(919) 677-8000
http://www.sas.com, software@sas.com

The SAS System is described as an integrated suite of software for enterprise-wide information delivery, which includes a major module for statistical analysis, as implemented in SAS/STAT. In release 8.1 two new experimental procedures (PROC MI and PROC MIANALYZE) were made available, which implemented multiple imputation. The interface for PROC MI changed substantially in release 8.2. SAS anticipates putting PROC MI and PROC MIANALYZE into production for release 9.

The imputation step is carried out by PROC MI, which allows use of either monotone (predictive mean matching, denoted by REGRESSION, or propensity, denoted by PROPENSITY) or nonmonotone (using MCMC) missingness methods. The MCMC methods can also be used in a hybrid model where the dataset is divided into monotone and nonmonotone parts, and a regression method is used for the monotone component. Extensive control and graphical diagnostics of the MCMC methods are provided. SAS supports transformation and back-transformation of variables. This may make an assumption of multivariate normality, needed for the REGRESSION and MCMC methods, more tenable (Schafer 1997).

Once PROC MI has been run, use of complete data methods is straightforward; the only addition is the specification of a BY statement to repeat these methods (i.e., PROC GLM, PROC PHREG, or PROC LOGISTIC) for each value of the variable `_Imputation_`. This approach is attractive, since it allows the full range of regression models available within SAS to be used in imputation.

The results are combined using PROC MIANALYZE, which provides a clear summary of the results. SAS provides an option (EDF) to use the adjusted degrees of freedom suggested by Barnard and Rubin (1999), and it displays estimates of the fraction of missing information for each parameter.

A disadvantage of the imputation methods provided by PROC MI is that the analyst has little control over the imputation model itself. In addition, for the regression and MCMC methods, SAS does not impute an observed value that is closest to the predicted value (i.e., there is no support for predictive mean matching using observed values). Instead, it uses an assumption of multivariate normality to generate a plausible value for the imputation. SAS allows the analyst to specify a minimum and maximum value for imputed values on a variable-by-variable basis, as well as the ability to round imputed values. In addition, a SAS data step could be used to generate observed values. In practice, however, both these approaches are somewhat cumbersome.

No additional installation was needed for PROC MI/PROC MIANALYZE, since they are bundled with SAS/STAT. The documentation was terse (69 pages for PROC MI, 31 pages for PROC MIANALYZE), but well organized; a number of examples were provided.

SAS is licensed on an annual basis, on a per-module basis. An annual license for base SAS and SAS/STAT is $3,900 for the first year, and $1,900 for subsequent years. Academic discounts are generally available.

Missing Data Library for S-Plus
Insightful (formerly MathSoft)
(800) 569-0123
http://www.insightful.com, sales@insightful.com

S-Plus 6.0 features a new missing data library, which extends S-Plus to support model-based missing data models, by use of the EM algorithm (Dempster, Laird, and Rubin 1977) and data augmentation (DA) algorithms (Tanner and Wong 1987). DA algorithms can be used to generate multiple imputations. The missing data library provides support for multivariate normal data (`impGauss`), categorical data (`impLoglin`), and conditional Gaussian models (`impCgm`) for imputations involving both discrete and continuous variables.

The package provides a concise summary of missing data distributions and patterns (further described in the discussion of the examples), including both text-based and graphical displays. There are good diagnostics provided for the convergence of the data augmentation algorithms. The printed documentation, while extensive (164 pages) is light on examples of imputation, instead focusing more on data augmentation and maximum likelihood (EM) based approaches. It provides an excellent tutorial regarding missing data methods in general.

S-Plus has strong support for file input, including the ability to connect directly to Excel, and to read files in a variety of formats (including Access, dBASE, Gauss, Matlab, Paradox, SAS, SPSS, Stata, and Systat).

The single-user license price for S-Plus 2000 Professional for Windows is $2500; S-Plus 6.0 is expected to have similar pricing. Discounted academic pricing is available including academic site licenses.

MICE
TNO Prevention and Health
Public Health
Wassenaarseweg 56
P.O. Box 2215
2301 CE Leiden
The Netherlands
(31) 71 518 18 18
http://www.multiple-imputation.com

Multiple Imputation by Chained Equations (MICE) is a library distributed for S-Plus (described above) and R, a system for statistical computation and graphics, whose language and interface is very similar to S-Plus.

MICE provides a variety of imputation models, including forms of predictive mean matching and regression methods, logistic and polytomous regression, and discriminant analysis. Nonmonotone missingness is handled by using chained equations (MCMC) to loop through all missing values. Extensive graphical summaries of the MCMC process are provided. In addition, MICE allows users to program their own imputation functions, which is useful for undertaking sensitivity analyses of different (possibly nonignorable) missingness models. The system allows transformation of variables, and fine-grained control over the choice of predictors in the imputation model. The imputation step is carried out using the `mice()` function. For continuous missing variables, MICE supports imputation using the `norm` function (similar to SAS' REGRESSION option), and the `pmm` function (similar to SOLAS' predictive mean matching). Completed datasets can be extracted using the `complete()` function, or can be run for each imputation using the `lm.mids()` or `glm.mids()` function. Finally, results can be combined using the `pool()` function.

Although computationally attractive, the chained equation approach implemented in MICE requires assumptions about the existence of the multivariate posterior distribution used for sampling, however, it is not always certain that such a distribution exists (van Buuren et al. 1999). Like SOLAS, MICE uses a fixed seed for random number generation, which must be overriden during the imputation phase to avoid always generating the same imputed values. It would be preferable to have this seed vary by default, but allow the option to fix the seed to allow replication of results.

Installation was straightforward, though automated addition of packages under R is only supported on Unix systems. In addition to the `mice` package, under R, two additional add-on packages were required (`MASS` and `nnet`). The documentation was short (39 pages), terse (particularly regarding the imputation models) but clear. An example using the NHANES dataset provided a summary of how to use the package. The manual included the help pages for each function in the library.

S-Plus was described previously. R is free software for Unix, Windows, and Macintosh that is distributed under a GNU-style copyleft. More information can be found at the R project web site: www.r-project.org. The MICE library is freely available, and may be downloaded from the www.multiple-imputation.com Web site.

Other Packages and Routines

Other packages that provide some support for imputation include SPSS, Joseph Schafer's free software (macros for S-Plus and stand-alone windows package NORM), Gary King's Amelia program, IVEware, HLM, and LISREL. Joseph Schafer's list of multiple imputation software routines (http://www.stat.psu.edu/~jls/misoftwa.html) and the list maintained by the Department of Statistics of TNO Prevention and Health, (http://www.multiple-imputation.com) are helpful in tracking the developments in this area.

SPSS requires the user to run the individual complete data models and combine the results. Schafer's software are an excellent companion to his book, but they do not support general purpose regression modeling. Amelia (http://gking.harvard.edu/stats.shtml) implements the EMis al-

```
> colSums(is.na(allison))
 x1 y    x2
 0  0    5008
> apply(allison, 2, anyMissing)
 x1 y x2
 F  F T
> round(100 * colMeans(is.na(allison)))
 x1 y  x2
 0  0  50
> M <- miss(allison); plot(M); M
Summary of missing values
     3 variables, 10000 observations, 2 patterns of missing values
        1 variables (33%) have at least one missing value
     5008 observations (50%) have at least one missing value
Breakdown by variable
 V O name  Missing % missing
 1 3    x2    5008        50
V = Variable number used below,  O = Original number (before sorting)
No missing values for variables: x1 y
```

Figure 2. Code to Describe Patterns of Missing Data for Artificial Data Example in S-Plus 6.0.

gorithm (King, Honaker, Joseph, and Scheve 2001) and performs the imputation step, but does not provide support for analysis of these imputed datasets or combination of the results. IVEware (http://www.isr.umich.edu/src/smp/ive) by Raghunathan et al. is a SAS version 6.12 callable routine built using the SAS macro language. It extends multiple imputation to support complex survey sample designs. HLM (hierarchical linear and nonlinear modeling) version 5 supports the analysis of multiply-imputed datasets, where multiple plausible datasets have been previously created. LISREL 8.20 and later supports multiple imputation, but the focus of this package is not on the regression models described in this article. Because of the existence of more complete implementations, these packages are not further discussed.

4. EXAMPLE: ARTIFICIAL DATA

We now replicate an artificial data example with missing covariates, as described by Allison (2000), to help illustrate how multiple imputation models are fit using these packages. Let X_1, X_2 and ϵ be multivariate normal with mean 0 and variance covariance given by:

$$\Sigma = \left\{ \begin{array}{ccc} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right\}.$$

The true regression model is given by:

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

We generated 10,000 observations from the multivariate normal distribution: $Y = X_1 + X_2 + \epsilon$ (i.e., $\beta_0 = 0$, $\beta_1 = \beta_2 = 1$). Following Allison's example, we caused approximately half of the values of X_2 in this dataset to be missing, according to the following mechanisms:

MCAR: X_2 is missing with probability 0.5
MAR $\mathbf{X}_1$: X_2 is missing if $X_1 < 0$
MAR Y: X_2 is missing if $Y < 0$
NI $\mathbf{X}_2$: X_2 is missing if $X_2 < 0$.

For approximately half the subjects $\mathbf{Z} = \mathbf{Z}^{\text{obs}} = (Y, X_1, X_2)$, while for the balance $\mathbf{Z}^{\text{obs}} = (Y, X_1)$ and $\mathbf{Z}^{\text{mis}} = X_2$. In this

Table 1. Parameter Estimates (and standard errors) From Artificial Datasets (true parameter values 1.00 and 1.00)

Missing mechanism	Parameter	Complete case	MI propensity	MI regression	MI MCMC
MCAR	X_1	1.00 (0.016)	1.40 (0.014)	1.01 (0.014)	1.01 (0.012)
	X_2	1.01 (0.016)	0.50 (0.016)	1.00 (0.013)	1.00 (0.015)
MAR X_1	X_1	1.04 (0.024)	1.54 (0.012)	1.02 (0.016)	1.01 (0.023)
	X_2	0.98 (0.016)	0.58 (0.019)	0.98 (0.014)	0.99 (0.015)
MAR Y	X_1	0.70 (0.015)	1.32 (0.013)	1.03 (0.014)	1.02 (0.015)
	X_2	0.71 (0.015)	0.83 (0.019)	1.00 (0.012)	1.00 (0.014)
NI X_2	X_1	1.00 (0.017)	1.28 (0.013)	1.15 (0.017)	1.16 (0.017)
	X_2	1.00 (0.025)	1.15 (0.027)	1.21 (0.026)	1.20 (0.020)

Table 2. Estimated Fraction of Missing Information for the MCAR Scenario for Two Replications (each with 10 imputations) of MICE PMM and SAS REGRESSION

Parameter	MICE 1	MICE 2	SAS 1	SAS 2
X_1	0.25	0.20	0.21	0.47
X_2	0.16	0.26	0.31	0.26

scenario, the complete case (CC) estimator (discarding all subjects with X_2 missing, the default option for missing data in most computing packages) is unbiased for the MCAR, MAR X_1, and NI X_2 mechanisms. For multiple imputation procedures (assuming MAR missingness), the MCAR, MAR X_1 and MAR Y missingness scenarios should yield unbiased estimates of the regression parameters. Allison (2000) showed that earlier versions of SOLAS (that provided only a propensity score method) had systematic severe bias when missingness depended on X_1, X_2, or Y.

Figure 2 displays the commands in the S-Plus 6.0 missing data library to describe patterns of missing data for the artificial data example. The other packages provided similar exploratory tools.

Table 1 displays the results from the imputation models using complete case, propensity (using SOLAS), predictive mean matching (REGRESSION, using SOLAS), and MCMC methods (using SAS) for each of the four missingness scenarios. The results, which are consistent with those of Allison, were similar when the imputation process was repeated on two occasions with 10 imputations each, or with one exception, when different programs (i.e., SOLAS, SAS, S-Plus 6.0, or MICE) were used. A bug (which was reported and corrected) in the propensity score routines in the beta release of SAS 8.2 yielded inaccurate variance estimates under the MAR Y scenario.

We note that for the MCAR missingness scenario, the multiple imputation estimators had slightly smaller standard errors than the complete case estimator. As expected, the complete case estimator was unbiased unless missingness depended on Y. When the MAR assumption was violated (NI X_2), imputation based methods were biased. In addition, the propensity score models were biased in all scenarios other than MCAR; this method is not recommended in this setting.

Table 2 displays estimates of the fraction of missing information for the MCAR scenario for two replications of MICE PMM and SAS REGRESSION. Both the parameters have more than 20% missing information. Note that even with 10 imputations, there remains a great deal of variability in the estimates of the fraction of missing information, though there was much less variability in the parameter estimates for the regression model in this example.

Figure 3 displays the code needed to fit the imputation model for each of the packages. The data are assumed stored in a dataset or object named `allison`.

```
                    SAS

proc mi data=allison out=miout nimpute=10 noprint;
        monotone method=reg;
        var y x1 x2;
proc reg data=miout outest=outreg covout noprint;
        model y = x1 x2;
        by _Imputation_;
proc mianalyze data=outreg;
        var Intercept x1 x2;
run;

                    S-plus 6.0 Missing Data library

library(missing)
emstart <- emGauss(allison)
worstFraction(emstart)
start <- matrix(rep(emstart$paramIter[2,],10),nrow=10,byrow=T)
imp <- impGauss(allison,start=start,control=list(niter=200))
fit <- miEval(lm(y ~ x1 + x2,data=imp))
result <- miMeanSE(fit)


                    MICE

library(mice)
imp <- mice(allison,imputationMethod="pmm",m=10,seed=456)
fit <- lm.mids(y ~ x1 + x2, imp)
result <- pool(fit)
```

Figure 3. Code to Fit Models for Artificial Data Example.

5. EXAMPLE: CHILD PSYCHOPATHOLOGY

We now consider an example with a missing outcome variable, from a study of child psychopathology in urban and rural Connecticut (Zahner, Jacobs, Freeman, and Trainor 1993; Zahner, Pawelkiewicz, DeFrancesco, and Adnopoz 1992; Zahner and Daskalakis 1997; Goldwasser and Fitzmaurice 2001). The measure of psychopathology used in the study was the internalizing problems scale (TXINT) of the Teacher's Report Form (Achenbach 1991b, TRF). TXINT will be used as the outcome for a linear regression model. TXINT ranges from 33 to 93 and can be considered to be approximately normal (Goldwasser and Fitzmaurice 2001).

In the study, 43% of teacher ratings on children were unobserved. Missingness of this magnitude is not uncommon: a similar rate was reported by Boyle et al. (1993) in their Ontario Child Health Study. There were a variety of causes of missingness for the teacher reports, including school district nonparticipation, parental refusal to give consent, and teacher nonresponse. Fitzmaurice, Laird, and Zahner (1996) considered the question of whether the missingness in this dataset is related to the unobserved teacher's rating, and found no evidence for this hypothesis. Thus, the missing at random assumption appears to be reasonable.

In addition to the teacher reports of psychopathology, parents also reported on the child, using a parallel instrument. Such multiple informant reports are commonly collected in services research, particularly for measurements of child psychopathology (Fitzmaurice, Laird, Zahner, and Daskalakis 1995; Offord et al. 1996; Horton and Laird 1999). We denote this report, based on the Child Behavioral Checklist (Achenbach 1991a, CBCL) as PXINT. We include in our analysis the 2,501 children with complete data on parent reports (of which 1,428 had complete teacher reports). For questions where primary interest involves the teacher report, this auxiliary information is extraneous to the regression model, but it may be associated with the unobserved teacher report. We will consider multiple imputation models that incorporate this auxiliary information.

Predictors of psychopathology, reported by the parent, included gender of the child (BOY: 1 = boy, 0 = otherwise); age of the child (OLD: 0 = age 6 to 8, 1 = age 9–11); social class (HIGH, MIDDLE, or LOW); area (RURAL, SUBURB, SMALL city, or LARGE city); maternal distress (MOMSTRS); child's health (HLTHPRO: 0 = good health, 1 = poor health); grade repetition (ACADPRO: 0 = no, 1 = yes); family stress (FAMSTRS: 0 = no, 1 = yes); and belonging to a single parent household (MOMSING: 0=father figure present, 1 = no father figure present).

Here we consider TXINT to be the outcome Y, which is not fully observed. X is fully observed, and consists of the above predictors. PXINT is auxiliary information that may improve the estimation of the missing values of TXINT.

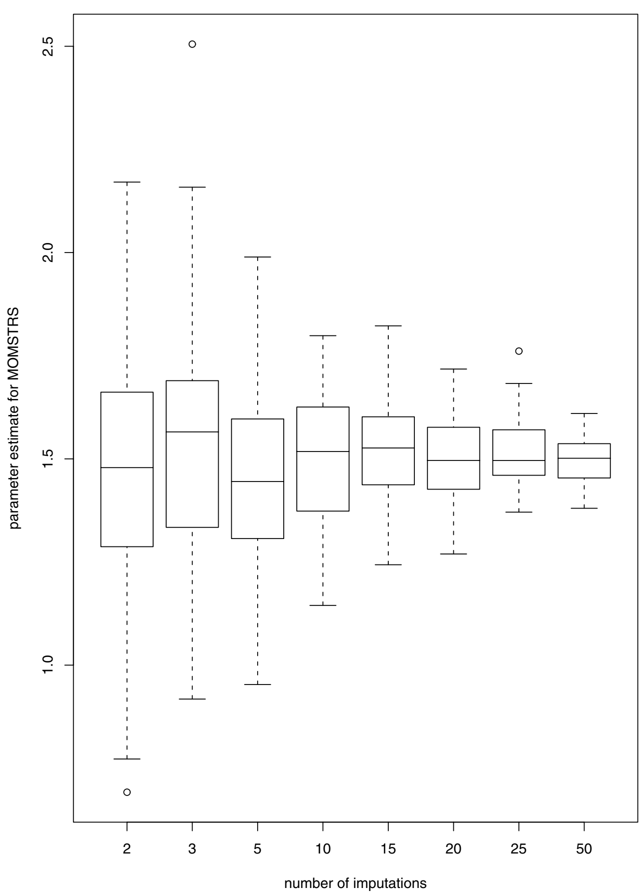

Figure 4. Distribution of Estimates of MOMSTRS Parameter Using Different Number of Imputations (based on 50 replications).

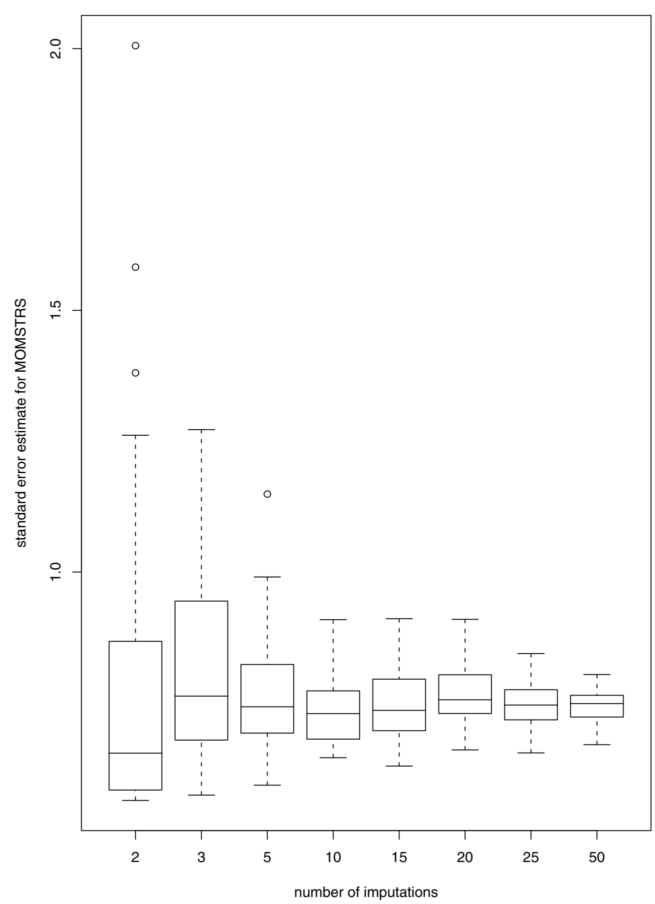

Figure 5. Distribution of Standard Error Estimates for MOMSTRS Parameter Using Different Number of Imputations (based on 50 replications).

In this setting (ordinary linear regression with fully observed discrete predictors plus a continuous auxiliary variable), multiple imputation is straightforward, due to the monotone nature of the missingness. We fit three models: one using just the complete cases, one that used multiple imputation, but not using the auxiliary information, and a model using multiple imputation with the auxiliary information (PXINT).

In this example, we noticed that with a small number of imputations, some of the parameter estimates were not stable when we repeated the imputation process with different starting seed values. We explored the variability of multiple imputation for this example by conducting a small simulation study. We repeated sets of imputations 50 times for each of the following numbers of imputations (2,3,5,10,15,20,25,50), and described the variability of the estimates of the MOMSTRS parameter. Figures 4 and 5 display boxplots of the results for the MOMSTRS parameter and the estimated standard error for this parameter. The median value for each set of imputations is approximately 1.5, but the variability of the results for a particular imputation is quite high. There is also increased variability in estimates of the standard error for this parameter. For the simulations with five imputations, the empirical 95% confidence interval (CI) for the MOMSTRS parameter from the 50 repetitions was (1.12,1.88). The size of this interval is similar to the magnitude of the standard error of this parameter, and could potentially affect the interpretation of the results. Even when 10 imputations were used, the empirical 95% CI was still large: (1.26,1.73). In contrast, when 50 imputations were conducted, the empirical 95% CI was smaller (1.40,1.60).

The variability of these estimates suggest that more imputations be used to decrease the variability in these models. This was true despite the fact that the fraction of missing information for this parameter was estimated to be in the range of 0.40.

Table 3 displays the parameter estimates from these three models for 50 imputations, using predictive mean matching in SAS. In general, the parameter estimates are roughly consistent in direction and magnitude, with similar estimates of standard errors. Other packages yielded similar results, though SOLAS allows a maximum of 10 imputations.

Despite the large number of imputations, the time involved in fitting the multiple imputation models for this example was not prohibitive. SAS, SOLAS, and MICE under S-Plus were run on a Dell Dimension XPS B866, while MICE under R was run on a 400Mhz Sun workstation. Five imputations were used in the timing studies, though the timing was approximately linear in the number of imputations. The models in SAS ran in 2 seconds (regression), 3 seconds (propensity), and 8 seconds (MCMC). The S-Plus missing data library using `impGauss()` required 13 seconds. The models in MICE under S-Plus ran in 13 seconds (regression), and 23 seconds (PMM). MICE under R was slightly slower (11 seconds and 82 seconds, respectively). These results may be due to the use of interpreted code under R, which was compiled to optimize certain functions under S-Plus, and the use of different hardware platforms and operating systems. SOLAS timing was not comparable; though the script language automates the imputation process, fitting the regression model required manual intervention by the analyst. This was unpleasant for the simulations described in this article, but should not be a factor in more standard settings. The predictive mean matching imputation model required 15 seconds, and approximately 1 minute to describe and fit the regression models and combine the results.

Table 3. Parameter Estimates (and standard error) for Child Psychopathology Example Using Complete Case Analysis and Multiple Imputation (using predictive mean matching, 50 imputations)

Parameter	*CC model*	*MI (no auxiliary)*	*MI (auxiliary)*
INTERCEPT	46.95 (0.80)	46.92 (0.81)	47.08 (0.82)
LARGE	0.75 (0.73)	0.87 (0.74)	0.74 (0.71)
SMALL	−1.51 (0.84)	−1.39 (0.83)	−1.58 (0.83)
SUBURB	0.65 (0.84)	0.65 (0.85)	0.47 (0.82)
LOW	2.98 (0.97)	3.07 (0.93)	3.05 (0.98)
MIDDLE	0.81 (0.60)	0.85 (0.59)	0.83 (0.57)
MOMSING	−0.31 (0.79)	−0.35 (0.81)	−0.48 (0.81)
MOMSTRS	1.57 (0.74)	1.60 (0.76)	1.58 (0.74)
HLTHPRO	0.67 (0.54)	0.65 (0.53)	0.58 (0.53)
ACADPRO	3.40 (0.57)	3.41 (0.62)	3.34 (0.57)
CSEX	−0.07 (0.54)	−0.11 (0.54)	−0.07 (0.53)
FAMSTRS	0.55 (0.56)	0.50 (0.58)	0.56 (0.55)

While these times are significantly slower than complete case methods, they are not prohibitive, even in a dataset of this size (2,501 observations with 13 variables). We concur with Rubin (1996) that computational time to carry about multiple imputation is no longer a serious concern for all but the largest datasets.

6. DISCUSSION

Earlier, we introduced common concerns associated with missing data: (1) loss of efficiency; (2) complication in data handling and analysis; and (3) bias due to differences between the observed and unobserved data. Multiple imputation is an analytic approach that addresses these problems. Compared to the complete case (CC) estimator, which discards partially observed subjects, multiple imputation methods may be more efficient (at the cost of making assumptions regarding the missingness distribution and the imputation model). Complications in data handling and analysis have been greatly simplified by the existence of easy-to-use software packages. If the MAR assumption is tenable, then multiple imputation may also provide less bias than other approaches if the imputation model is correctly specified. Although beyond the scope of this article, additional research is needed to investigate the bias resulting from a poorly specified imputation model (e.g., using a normal distribution when the possibly missing variable is Bernoulli).

The previously described packages for multiple imputation, though still evolving, are a useful addition to the analytic toolchest of practicing statisticians. They facilitate use of multiple imputation without having to resort to custom programming, tedious housekeeping, and additional calculations. Overall, these packages provide an easy to use environment for imputation, and allow this technique to be applied in many settings.

None of the packages is clearly superior, and there is a great deal of overlap of their support for multiple imputation. SOLAS provides a nice interface in a package that is geared primarily toward imputation, but as a special-purpose package, it is limited in its general purpose statistical coverage. It is fine for linear regression but does not currently facilitate imputation analysis for nonlinear models. SAS provides a general purpose environment for imputation, but does not provide as fine-grained control of the imputation model. The S-Plus missing data library and MICE are somewhat slower than the other two packages, but provide support for a wide variety of regression models.

Extensive MCMC diagnostics are provided by both SAS and MICE. Although these methods are quite attractive in modeling nonmonotone missingness, they remain in large part a complicated black box whose output can be difficult to interpret. We note that because both of the examples used to illustrate the methods featured a monotone missing data pattern, these routines were not fully explored in this review.

All of the packages allow the number of imputations to be varied (though SOLAS limits this to 10). Since this number is under the control of the analyst, and because it affects the variability of results, it is important that the sensitivity of the results to the number of imputations be explored. Given the efficiency and speed of current computers, use of more than 10 imputations can easily be incorporated into most analyses.

The existing packages all allow the incorporation of auxiliary information, which may improve the estimation of the imputation model. Such models are often useful as an adjunct to a complete case analysis, as an informal test of sensitivity to assumptions required by complete data methods. These packages (with the exception of MICE) do not allow the use of sensitivity analysis of nonignorable nonresponse. Such models, as described in detail by Rubin (1987, chap. 6), assess the robustness of inference to different assumptions about missingness. As various authors have noted, sensitivity analyses are quite complicated, but incorporation of simple nonignorable missingness models would be a useful addition to these packages. MICE is the only package that allows the analyst to construct their own imputation functions, implementing particular nonignorable assumptions, but these new functions require additional programming.

As a closing note, we offer the reminder that multiple imputation is not a panacea. Although it is a powerful and useful tool applicable to many missing data settings, if not used carefully it is potentially dangerous. The existence of software that facilitates its use requires the analyst to be careful about the verification of assumptions, the robustness of imputation models, and the appropriateness of inferences. For more complicated models (e.g., longitudinal or clustered data), this is even more important. We reaffirm the sage advice of Barnard and Meng (1999):

> Cautions are needed, however, just as with any statistical methodology. It is clear that if the imputation model is seriously flawed in terms of capturing the missing-data mechanism, then so will be any analysis based on such imputations. This problem can be avoided by carefully investigating each specific application, by making the best use of knowledge and data about the missing-data mechanism, and by performing various model checking procedures, in particular, posterior predictive checks. This is not an additional burden for using Rubin's method, but rather a fundamental requirement for any general method that attempts to produce statistically and scientifically meaningful results in the presence of incomplete data.

[Received April 2001. Revised May 2001.]

REFERENCES

Achenbach, T. M. (1991a), *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*, University of Vermont: Department of Psychiatry.

——— (1991b), *Manual for the Teacher's Report Form and 1991 Profile*, University of Vermont: Department of Psychiatry.

Allison, P. D. (2000), "Multiple Imputation for Missing Data: A Cautionary Tale," *Sociological Methods and Research*, 28, 301–309.

Barnard, J., and Meng, X. L. (1999), "Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES," *Statistical Methods in Medical Research*, 8, 17–36.

Barnard, J., and Rubin, D. B. (1999), "Small Sample Degrees of Freedom With Multiple Imputation," *Biometrika*, 86, 948–955.

Boyle, M. H., Offord, D. R., Racine, Y. A., Fleming, J. E., Szatmari, P., and Links, P. S. (1993), "Predicting Substance Use in Early Adolescence Based on Parent and Teacher Assessments of Childhood Psychiatric Disorder: Results From the Ontario Child Health Study Follow-up," *Journal of Child Psychology and Psychiatry*, 34, 535–544.

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.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford: Clarendon Press.

Fienberg, S. E. (1994), "Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality," *Journal of Official Statistics*, 10, 115–132.

Fienberg, S. E., and Willenborg, L. C. R. J. (1998), "Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data," *Journal of Official Statistics*, 14, 337–345.

Fitzmaurice, G. M., Laird, N. M., and Zahner, G. E. P. (1996), "Multivariate Logistic Models for Incomplete Binary Responses," *Journal of the American Statistical Association*, 91, 99–108.

Fitzmaurice, G. M., Laird, N. M., Zahner, G. E. P., and Daskalakis, C. (1995), "Bivariate Logistic Regression Analysis of Childhood Psychopathology Ratings Using Multiple Informants," *American Journal of Epidemiology*, 142, 1194–1203.

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

Goldwasser, M. A., and Fitzmaurice, G. M. (2001), "Multivariate Linear Regression of Childhood Psychopathology using Multiple Informant Data," *International Journal of Methods in Psychiatric Research*, 10.

Horton, N. J., and Laird, N. M. (1999), "Maximum Likelihood Analysis of Generalized Linear Models with Missing Covariates," *Statistical Methods in Medical Research*, 8, 37–50.

Ibrahim, J. G. (1990), "Incomplete Data in Generalized Linear Models," *Journal of the American Statistical Association*, 85, 765–769.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001), "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review*, 95, 49–69.

Laird, N. M. (1988), "Missing Data in Longitudinal Studies," *Statistics in Medicine*, 7, 305–315.

Little, R. J. A. (1988), "Missing-Data Adjustments in Large Surveys," *Journal of Business and Economic Statistics*, 6, 287–296.

——— (1992), "Regression With Missing X's: A Review," *Journal of the American Statistical Association*, 87, 1227–1237.

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

Liu, M., Taylor, J. M. G., and Belin, T. R. (2000), "Multiple Imputation and Posterior Simulation for Multivariate Missing Data in Longitudinal Studies," *Biometrics*, 56, 1157–1163.

Meng, X.-L., and Rubin, D. B. (1992), "Performing Likelihood Ratio Tests With Multiply-Imputed Datasets," *Biometrika*, 79, 103–111.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., and Lipman, E. L. (1996), "Integrating Assessment Data from Multiple Informants," *Journal of the American Academy of Child Adolescent Psychiatry*, 35, 1078–1085.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors are not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

——— (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

Robins, J. M., and Wang, N. (2000), "Inference for Imputation Estimators," *Biometrika*, 87, 113–124.

Rosenbaum, P. R., and Rubin, D. B. (1983), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Series B, 45, 212–218.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–590.

——— (1977), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

——— (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91, 473–489.

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

——— (1999), "Multiple Imputation: A Primer," *Statistical Methods in Medical Research*, 8, 3–15.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.

van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999), "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis," *Statistics in Medicine*, 18, 681–694.

Wang, N., and Robins, J. M. (1998), "Large-Sample Theory for Parametric Multiple Imputation Procedures," *Biometrika*, 85, 935–948.

Zahner, G. E. P., and Daskalakis, C. (1997), "Factors Associated with Mental Health, General Health and School-Based Service use for Psychopathology," *American Journal of Public Health*, 87, 1440–1448.

Zahner, G. E. P., Jacobs, J. H., Freeman, D. H., and Trainor, K. (1993), "Rural-Urban Child Psychopathology in a Northeastern U.S. State: 1986–1989," *Journal of the American Academy of Child Adolescent Psychiatry*, 32, 378–387.

Zahner, G. E. P., Pawelkiewicz, W., DeFrancesco, J. J., and Adnopoz, J. (1992), "Children's Mental Health Service Needs and Utilization Patterns in an Urban Community," *Journal of the American Academy of Child Adolescent Psychiatry*, 31, 951–960.

Zaslavsky, A. M., and Horton, N. J. (1998), "Balancing Disclosure Risk Against the Loss of Nonpublication," *Journal of Official Statistics*, 14, 411–419.