Multiple imputation for multivariate missing-data problems: a data analyst's perspective

Joseph L. Schafer and Maren K. Olsen

The Pennsylvania State University

March 9, 1998

Abstract

Analyses of multivariate data are frequently hampered by missing values. Until recently, the only missing-data methods available to most data analysts have been relatively ad hoc practices such as listwise deletion. Recent dramatic advances in theoretical and computational statistics, however, have produced a new generation of flexible procedures with a sound statistical basis. These procedures involve multiple imputation (Rubin, 1987), a simulation technique that replaces each missing datum with a set of $m > 1$ plausible values. The m versions of the complete data are analyzed by standard complete-data methods, and the results are combined using simple rules to yield estimates, standard errors, and p-values that formally incorporate missing-data uncertainty. New computational algorithms and software described in a recent book (Schafer, 1997) allow us to create proper multiple imputations in complex multivariate settings. This article reviews the key ideas of multiple imputation, discusses the software programs currently available, and demonstrates their use on data from the Adolescent Alcohol Prevention Trial (Hansen & Graham, 1991).

1 Introduction

Missing data are a familiar problem to researchers in the social and behavioral sciences. In longitudinal studies, subjects may drop out early or be unavailable during one or more data collection periods. When data are collected by questionnaire, subjects may be unwilling or unable to respond to some items, or may fail to complete sections of the questionnaire due to lack of time or interest. These types of missingness, though inevitable, are unintended and uncontrolled by the researcher. For some data collection efforts, it may also be desirable to incorporate planned missingness into the study design. For example, the Adolescent Alcohol Prevention Trial (Hansen & Graham, 1991) involved a three-form design in which less essential items were divided into three sections, A, B, and C; one-third of the subjects received sections A and B, one-third received B and C, and one-third received A and C. This design allowed researchers to measure relationships among a greater number of items without increasing the burden on individual respondents.

Until recently, the only methods widely available for analyzing incomplete data focused on "removing" the missing values, either by ignoring subjects with incomplete information or by substituting plausible values (e.g. means or regression predictions) for the missing items. These ad hoc methods, though simple to implement, have serious drawbacks which have been well documented (see, e.g. Little & Rubin, 1987; Graham, Hofer, & Piccinin, 1994; and their references). For multivariate analyses involving a large number of items, case deletion procedures can be very inefficient, discarding an unacceptably high proportion of subjects; even if the per-item rates of missingness are low, few subjects may have complete data for all items. In addition, case-deletion procedures may bias the results if the subjects who provide complete data are unrepresentative of the entire sample. Simple

mean substitution will seriously dampen relationships among variables, but substituting regression predictions will artificially inflate correlations. Even if the missing values could be imputed in such a way that the distributions of variables and relationships among them were perfectly preserved, the imputed dataset would still fail to provide accurate measures of variability for the following reason: subsequent analyses would fail to account for missing-data uncertainty. Regardless of the imputation method, imputed values are only estimates of the unknown true values. Any analysis that ignores the uncertainty of missing-data prediction will lead to standard errors that are too small, p-values that are artificially low, and rates of Type I error that are higher than nominal levels.

Fortunately, in the last two decades, substantial progress has been made in developing statistical procedures for missing data. In the late 1970's, Dempster, Laird, & Rubin (1977) formalized the EM algorithm, a computational method for efficient estimation from incomplete data. EM has proven to be very useful as a computational device. More importantly, the ideas underlying EM signalled a fundamental shift in the way statisticians viewed missing data. Until that time, missing data were viewed as a nuisance *to be gotten rid of*, either by case deletion or imputation. Since then, statisticians have begun to see missing values as a source of variability *to be averaged over*. In any incomplete dataset, the observed values provide indirect evidence about the likely values of the unobserved ones. This evidence, when combined with certain assumptions (described below), implies a predictive probability distribution for the missing values that should be averaged over in the statistical analysis. Modern missing-data techniques carry out this averaging in a variety of ways. EM algorithms perform the averaging in a deterministic or non-random fashion. More recently, Rubin (1987) has developed the paradigm of *multiple imputation*, which carries out the averaging via simulation.

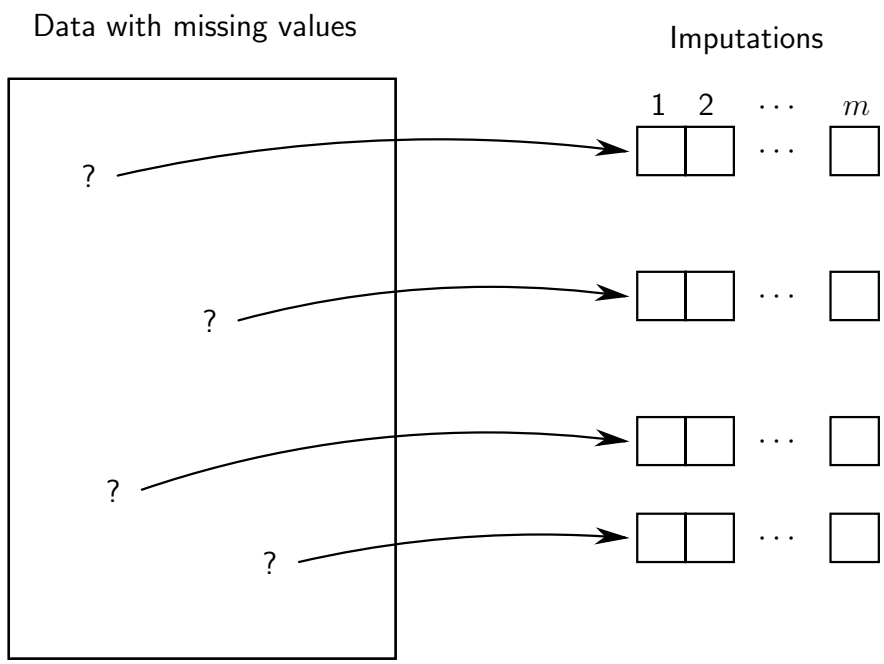


Figure 1: Matrix of multivariate data with missing values and multiple imputations

In multiple imputation (MI), each missing value is replaced by a set of $m > 1$ plausible values drawn from their predictive distribution. A multiply imputed dataset is depicted in Figure 1. The variation among the m imputations reflects the uncertainty with which the missing values can be predicted from the observed ones. After performing MI there are m apparently complete datasets, each of which can be analyzed by complete-data methods. After performing identical analyses on each of the m datasets, the results (estimates and standard errors) are combined, using simple rules provided by Rubin (1987) and others, to produce overall estimates and standard errors that reflect missing-data uncertainty.

MI is attractive for a number of reasons. First, it works in conjunction with standard complete-data methods and software. One the MI's have been generated, the analyses can be carried out using procedures in SAS, LISREL, or virtually any other statistical package. Second, one set of m imputations may be used for a variety of analyses; there is often no

Table 1: Percent efficiency of MI estimation by number of imputations m and fraction of missing information γ

	γ				
m	.1	.3	.5	.7	.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

need to reimpute when a new analysis is performed. Third, the inferences—standard errors, p-values, etc.—obtained from MI are generally valid because they incorporate uncertainty due to missing data. Finally, MI is attractive because it can be highly efficient even for small values of m. In many applications, just 3–5 imputations are sufficient to obtain excellent results.

Many are surprised by the claim that only 3–5 imputations may be needed. Rubin (1987, p. 114) shows that the efficiency of an estimate based on m imputations is approximately

$$\left(1 + \frac{\gamma}{m}\right)^{-1},$$

where γ is the *fraction of missing information* for the quantity being estimated. The fraction γ quantifies how much more precise the estimate might have been if no data had been missing. (A generic method for estimating γ will be given below.) The efficiencies achieved for various values of m and rates of missing information are shown in Table 1. This table shows that gains rapidly diminish after the first few imputations. Consider the column for

30% missing information ($\gamma = .3$), a moderately high rate for many applications. With $m = 5$ imputations, we have already achieved 94% efficiency. Increasing the number to $m = 10$ raises the efficiency to 97%, a rather slight gain for a doubling of computational effort. In most situations there is simply little advantage to producing and analyzing more than a few imputed datasets.

Although MI was first proposed nearly twenty years ago (Rubin, 1978), the method has remained largely unknown and unused by non-experts. The main reason for this obscurity has been the lack of computational tools for creating MI's. Except in trivial settings, the probability distributions that one must draw from to produce proper MI's tend to be complicated and intractable. Very recently, however, a remarkable variety of new simulation methods have appeared in the statistical literature. These methods, known collectively as *Markov chain Monte Carlo*, have spawned a small revolution in applied parametric modeling (Gilks, Richardson, & Spiegelhalter, 1996). Schafer (1997) has adapted and implemented Markov chain Monte Carlo methods for the purpose of multiple imputation. In particular, he has written general-purpose MI software for incomplete multivariate data. Some of these programs, which are easier to learn and use than previous ones, are implemented as graphical, stand-alone applications for PCs running Windows (95/NT). They may be downloaded free of charge at our web site (`http://stat.psu.edu/~jls/misoftwa.html`). Four packages are currently available: NORM, which performs multiple imputation under a multivariate normal model; CAT, for multivariate categorical data; MIX, for mixed datasets containing both continuous and categorical variables; and PAN, for multivariate panel or clustered data.

In the remainder of this article, we will introduce and explain the key ideas of MI and the use of our software from the perspective of a data analyst. First we review the major assumptions that are made about the data and mechanisms of nonresponse. Then we

provide an overview of some computational tools and algorithms used by Schafer's software, and summarize Rubin's rules for combining point estimates and standard errors from multiple analyses. The latter part of this article illustrates the use of MI on data drawn from the Adolescent Alcohol Prevention Trial (Hansen & Graham, 1991), which we impute and analyze using NORM. Finally, we conclude with a discussion comparing MI to some other modern missing-data procedures currently available, such as those used by the structural equations modeling program Amos (Arbuckle, 1995).

2 Assumptions

Like any statistical method, MI is based upon certain assumptions. Responsible use of MI requires basic understanding of these assumptions and the possible implications for subsequent analyses if they are violated. These assumptions pertain to (a) the population of data values, (b) the prior distribution for the model parameters, and (c) the mechanism of nonresponse. Let us consider each of these topics in turn.

2.1 The data model

In order to generate imputations for the missing values, one must impose a probability model on the complete data (observed and missing values). Each of our software packages applies a different class of multivariate complete-data models. NORM uses the multivariate normal distribution. CAT is based on loglinear models, which have been traditionally used by social scientists to describe associations among variables in cross-classified data. The MIX program relies on the *general location model*, which combines a loglinear model for the

categorical variables with a multivariate normal regression for the continuous ones. Details of these models are given by Schafer (1997). The new package PAN, described by Schafer (submitted), uses a multivariate extension of a popular two-level linear regression model commonly applied to multilevel data (e.g. Bryk & Raudenbush, 1992). The PAN model is appropriate for describing multiple variables collected on a sample of individuals over time, or multiple variables collected on individuals who are grouped together into larger units (e.g. students within classrooms).

Experienced analysts know that real data rarely conform to convenient models such as the multivariate normal. In most applications of MI, the model used to generate the imputations will at best be only approximately true. Fortunately, experience has repeatedly shown that MI tends to be quite forgiving of departures from the imputation model. For example, when working with binary or ordered categorical variables, it is often acceptable to impute under a normality assumption and then round off the continous imputed values to the nearest category. Variables whose distributions are heavily skewed may be transformed (e.g. by taking logarithms) to approximate normality and then transformed back to their original scale after imputation. Simulation studies demonstrating the robustness of MI to departures from the imputation model are reported by Ezzati-Rice et al. (1995); Schafer (1997); and Graham & Schafer (in press).

Despite these encouraging results, it is naive to suggest that imputation may be carried out haphazardly or that the choice of imputation model is unimportant. An imputation model should be chosen to be (at least approximately) compatible with the analyses to be performed on the imputed datasets. In particular, the model should be rich enough to preserve the associations or relationships among variables that will be the focus of later investigation. For example, suppose that a variable Y is imputed under a normal model that

includes the variable X_1. After imputation, the analyst then uses linear regression to predict Y from X_1 and another variable X_2 which was not in the imputation model. The estimated coefficient for X_2 from this regression would tend to be biased toward zero, because Y has been imputed without regard for its possible relationship with X_2. In general, any association that may prove important in subsequent analyses should be present in the imputation model.

The converse of this rule, however, is not at all necessary. If Y has been imputed under a model that includes X_2, there is no need to include X_2 in future analyses involving Y unless its relationship to Y is of substantive interest. Results pertaining to Y cannot be biased by the inclusion of extra variables in the imputation phase. Therefore, a rich imputation model that preserves a large number of associations is desirable because it may be used for a variety of post-imputation analyses.

2.2 The prior distribution

The statistical theory underlying MI involves the fundamental law of probability known as Bayes's Theorem. The Bayesian nature of MI requires the imputer to specify a prior distribution for the parameters of the imputation model. In the Bayesian paradigm, this prior distribution quantifies one's belief or state of knowledge about model parameters before any data are seen. Because different prior distributions can lead to different results, Bayesian methods have been regarded by some as subjective and unscientific. In practice, however, the results of a Bayesian procedure tend to be far more sensitive to the choice of the data model than the choice of the prior. In many cases—especially when the sample size is moderately large—nearly any reasonable prior distribution should lead to essentially the same results.

We tend to view the prior distribution as a necessary evil, a mathematical convenience that allows us to generate the imputations in a principled fashion. By default, our software applies well accepted "noninformative" prior distributions that correspond to a state of prior ignorance about model parameters. In the vast majority of data analyses, the default noninformative prior works well. In some unusual situations—with small samples, sparse data or high rates of missing information—it may be necessary to apply an informative prior distribution. Extended discussion on the choice of prior distributions is given by Schafer (1997).

2.3 The nonresponse mechanism

Every missing-data method must make some largely untestable statistical assumptions about the manner in which the missing values were lost. Our methods assume that the missingness mechanism is *ignorable*, in the precise sense defined by Rubin (1987). Another, essentially equivalent, term for this is the *missing at random* (MAR) assumption (Rubin, 1976; Little & Rubin, 1987). Despite its name, MAR does not mean that the missing values must be a random subsample of the entire dataset. The latter condition, known as *missing completely at random*, is much more restrictive and often unrealistic. In layman's terms, MAR means that the probabilities of missingness may depend on data values that are observed but not on ones that are missing.

To understand the MAR assumption, consider a simple bivariate dataset with one variable X that is always observed and a second variable Y that is sometimes missing. Under MAR, the probability that Y is missing for a sample subject may be related to the subject's value of X but not to his value of Y. In other words, suppose we define a response indicator

R that is equal to one if Y is observed and zero if Y is missing; MAR implies that the Y and R may be related, but only indirectly through their mutual associations with X. Under MAR, the values of Y for nonrespondents may tend to be systematically higher or lower than those for respondents. But MAR does imply that that the statistical *relationship* (in the sense of regression) between Y and X is on average no different for the two groups. If MAR were satisfied, then one could regress Y on X for the respondents and then use the estimated relationship to obtain unbiased predictions of Y for nonrespondents.

Extensions of this simple example to more complex sitations—e.g. to where missing values occur on both X and Y—are less intuitive and more difficult to describe. Nevertheless, the prediction principle does carry over. MAR is the formal assumption that allows us to first estimate the relationships among variables from the observed data, and then use these relationships to obtain unbiased predictions of the missing values from the observed values.

Regarding MAR, several points ought to be made. First, MAR is defined relative to the variables present in a dataset. If a variable X is related both to the missingness of other variables and to the values of those variables, and X is removed from the dataset, then MAR will no longer be satisfied. For this reason, it is good practice to include in an imputation procedure variables that are predictive of missingness, because MAR then becomes more plausible. Second, the MAR hypothesis cannot be tested from data at hand; doing so would require knowledge of the missing values themselves. Third, data that are missing by design (e.g. in a study with planned missingness) are generally MAR.

When data are missing for reasons beyond the control of the investigators, one can never be sure whether MAR holds. In fact, to speak of a single "missingness mechanism" is often misleading, because in most studies missing values occur for a variety of reasons; some

of these reasons may be entirely unrelated to the data in question, but others may be closely related. Consider, for example, the problem of attrition in a school-based longitudinal study of adolescent substance use; over time, some subjects may be lost simply because they moved to another school district, whereas others may have dropped out of school in a pattern of problematic behavior that includes substance use. Uncontrolled missingness typically arises from a mixture of ignorable and nonignorable sources. For a detailed discussion of this point with examples of both planned and uncontrolled missingness, see Graham, Hofer, and Piccinin (1994).

Unfortunately, it is not possible to relax the MAR assumption in any meaningful way without replacing it with other equally untestable assumptions. The alternative to assuming MAR is to propose a formal probability model for response and carry out an analysis under that model. Doing so usually requires substantial effort and technical expertise. Such analyses have been done in a variety of settings, but the methods tend to be rather problem-specific and do not generalize well. At present, there are no principled nonignorable missing-data methods readily available to most data analysts. Until such methods become available, we recommend the cautious use of ignorable methods with an awareness of their limitations. In the vast majority of studies, principled methods that assume MAR will tend to perform better than ad hoc procedures such as listwise deletion or imputation of means. Moreover, even when MAR seems unrealistic, ignorable procedures that draw upon a rich assortment of covariates may introduce little or no bias relative to a complicated nonignorable procedure. Further discussion of nonignorable alternatives is given by Schafer (1997, pp. 27–28).

3 Tools for creating multiple imputations

Special computational techniques are needed to create MI's for incomplete multivariate data. Our software applies new methods of Markov chain Monte Carlo. In particular, we use a procedure called *data augmentation* (Tanner & Wong, 1987). Because data augmentation is closely related to the EM algorithm (Dempster, Laird, and Rubin, 1977), we first review EM.

3.1 The EM algorithm

The EM algorithm is a general technique for fitting models to incomplete data. EM capitalizes on the relationship between missing data and the unknown parameters of a data model. If we knew the missing values, then estimating the model parameters would be straightforward. Similarly, if we knew the parameters of the data model, then it would be possible to obtain unbiased predictions for the missing values. This interdependence between model parameters and missing values suggests an iterative method where we first predict the missing values based on assumed values for the parameters, use these predictions to update the parameter estimates, and repeat. The sequence of parameters converges to maximum-likelihood estimates that implicitly average over the distribution of the missing values.

The formal definition and key properties of EM are reviewed by Little & Rubin (1987) and Schafer (1997). It is helpful to note that EM's rate of convergence is determined by the rates of missing information in the dataset. If there were no missing values, then convergence would be immediate; if large amounts of information are missing about one or

more parameters, then convergence will require many iterations. One way to monitor the convergence of EM is to examine the loglikelihood function and confirm that it increases at each iteration. In addition, it is sometimes helpful to run EM from a few different starting values to ensure that the loglikelihood has a unique maximum. There are situations where a unique maximum does not exist; for example, the likelihood might have multiple modes or a ridge. These situations tend to occur with small samples, high rates of missingness, and models that have too many parameters to be supported by the amount of observed data.

The programs NORM, CAT, and MIX include EM-type algorithms for parameter estimation. Running these algorithms before producing MI's is highly recommended for two reasons: (a) the resulting parameter estimates provide excellent starting values for the data augmentation procedure, and (b) the convergence behavior of EM helps to predict the convergence behavior of subsequent data augmentation runs.

3.2 Data augmentation

Like EM, data augmentation (DA) is an iterative process that alternately fills in the missing data and makes inferences about the unknown parameters. However, DA does this in a stochastic or random fashion. DA first performs a random imputation of missing data under assumed values of the parameters, and then draws new parameters from a Bayesian posterior distribution based on the observed and imputed data. The procedure of alternately simulating missing data and parameters creates a Markov chain that eventually stabilizes or converges in distribution. The distribution of the parameters stabilizes to a posterior distribution that averages over the missing data. The distribution of the missing data stabilizes to a predictive distribution—the exact distribution, in fact, that one needs to draw from to

create proper MI's.

As with EM, the convergence of DA is closely related to rates of missing information; large amounts of missing information cause the convergence to be slow. The meaning of convergence, however, is quite different because DA is a stochastic procedure that converges *in distribution*. When EM converges, the parameter estimates no longer change from one iteration to the next. When DA converges, the distribution of the parameters no longer changes from one iteration to the next, although the random parameter values themselves do continue to change. For this reason, assessing the convergence of DA is much more complicated than for EM.

One way to address this issue is to reinterpret convergence as a lack of serial dependence. DA may be said to have converged by k cycles if the value of any parameter at cycle t is statistically independent of its value at cycle $t + k$ for $t = 1, 2, \ldots$. We can assess the degree of serial dependence by storing the parameters at each cycle and constructing time-series plots. Storing the parameter values also allows us to calculate and plot the sample autocorrelation function (ACF) for any parameter. The ACF is simply the lag-k Pearson correlation coefficient for various values of k, i.e. the correlation between the simulated value of a parameter at any cycle and its value k cycles later. A DA algorithm can reasonably be said to have converged by k cycles if the sample ACFs for all parameters have died down to zero by lag k. Our software is capable of storing parameters and creating time-series and ACF plots automatically, making it easy for the user to monitor convergence.

In practice, we have found that DA nearly always converges in fewer cycles than does EM. For example, if the EM algorithm for a particular problem converges in 50 cycles, it would be highly unusual for the corresponding DA algorithm to require more than 50 cycles.

For this reason, it is very helpful to run the EM algorithm prior to running DA to obtain a rough idea of how many DA cycles may be needed.

When using DA to create m multiple imputations of missing data, it is important to note that proper MI's are *independent* draws of missing data from their predictive distribution. Once we have determined that DA has converged by k cycles, we can perform m runs of length k and save the completed datasets from the end of each run as our m imputations. Alternatively, we can perform a single run of length km and store the completed datasets from cycles $k, 2k, \ldots, mk$. Our software is designed to perform either operation. The user must specify starting values for the parameters (typically the estimates obtained from EM) and a total run length. The user may then optionally choose to save the final imputed datasets from the end of the DA run, or to save the imputed datasets at every kth cycle for any value of k.

There is no danger in using too large a value of k, because once DA has converged the process remains stationary. On the other hand, if k is too small, the imputed datasets will not be truly independent and the variability among them may understate the true levels of missing-data uncertainty. Time and computational resources allowing, it is therefore advisable to choose k to be somewhat larger than necessary, to create a margin of safety and ensure that the multiple imputations are truly independent.

4 Rules for MI inference

Once MI's have been created, the datasets may be analyzed by nearly any method that would be appropriate if the data were complete. For example, one could perform linear

or logistic regression procedures using any standard statistical package. Any model would have to be fit m times, once for each imputed dataset, and the results across these datasets will vary as a reflection of missing-data uncertainty. To obtain an overall set of estimated coefficients and standard errors, one would have to store the estimates and standard errors from each of the m imputed datasets, and then combine the results using the rules given by Rubin (1987). These same rules have been implemented in each of our software packages; each program (NORM, CAT, MIX, and PAN) has the capacity to read m files containing point estimates and standard errors and combine them by these rules. Alternatively, these rules could easily be programmed as a macro in SAS or another statistical package.

Rubin's rules proceed as follows. Let $\hat{Q}$ denote an estimate of a population quantity of interest and U its estimated variance. For example, $\hat{Q}$ could be an estimated regression coefficient and U its squared standard error. After performing the same analysis on each imputed dataset, we have m equally plausible estimates $\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_m$ and their corresponding variances $U_1, U_2, \ldots, U_m$. The MI estimate, or overall estimate, is given by

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i.$$

The total variance for the estimate has two components that take into account variability within each dataset and across datasets. The within-imputation variance,

$$\bar{U} = \frac{1}{m} \sum_{i=1}^{m} U_i,$$

is simply the average of the estimated variances. The between-imputation variance,

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{Q}_i - \bar{Q})^2,$$

is the sample variance of the estimates themselves. The total variance, T, is the sum of the two components with an additional correction factor to account for the simulation error in

$\bar{Q}$,

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B.$$

The square root of T is the overall standard error associated with $\bar{Q}$. Note that if there were no missing data, then $\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_m$ would be identical, B would be 0 and T would simply be equal to $\bar{U}$. The size of B relative to $\bar{U}$ is a reflection of how much information is contained in the missing part of the data relative to the observed part.

A rough 95% confidence interval can be obtained as $\bar{Q} \pm 2\sqrt{T}$. In general, however, it is better to calculate intervals using the approximation

$$\bar{Q} \pm t_{df}\sqrt{T},$$

where t_{df} denotes a quantile of Student's t-distribution with degrees of freedom

$$df = (m-1)\left(1 + \frac{m\bar{U}}{(m+1)B}\right)^2.$$

P-values for testing the null hypothesis $Q = 0$ may be obtained by comparing the ratio $\bar{Q}/\sqrt{T}$ to the same t-distribution.

Notice that with an infinite number of imputations ($m = \infty$) the total variance reduces to the sum of the two variance components and the confidence interval is based on a normal distribution ($df = \infty$). The degrees of freedom are influenced both by the number of imputations and the relative sizes of B and $\bar{U}$. When B dominates $\bar{U}$ the degrees of freedom are close to the minimum value of $m - 1$, but when $\bar{U}$ dominates B the degrees of freedom approach infinity. If the computed value of df is very small—say, less than 10—it suggests that greater efficiency (i.e. more accurate estimates and narrower intervals) could be obtained by increasing the number of imputations m. If df is large, however, it suggests that little will be gained from a larger m.

Rubin (1987) also shows that an estimate of the fraction of missing information about the population quantity Q is

$$\gamma = \frac{r + 2/(df + 3)}{r + 1},$$

where

$$r = \frac{(1 + m^{-1})B}{\bar{U}}$$

is the relative increase in variance due to nonresponse. Both of these quantities are interesting and useful diagnostics, revealing how strongly the estimation of Q may be influenced by missing data.

Many important statistical questions cannot be answered by a confidence interval or hypothesis test about a single parameter Q. One is often concerned with assessing the joint significance of group of coefficients in a model, or testing a model's overall goodness of fit. General rules for multiparameter inference from multiply-imputed data are reviewed by Schafer (1997, pp. 112–118). Unfortunately, the implemention of these rules with existing statistical software is not necessarily straightforward. We are currently investigating methods for goodness-of-fit testing and model selection in structural equation models using popular software packages such as LISREL; new results in this area should be available soon.

5 Application: The Adolescent Alcohol Prevention Trial

5.1 The data

The data in this example come from the Adolescent Alcohol Prevention Trial (AAPT), a school-based study of substance use in the Los Angeles area (Hansen & Graham, 1991).

We examine a cohort of $N = 3,017$ students who received an intervention in grade 7. Students were assigned to four groups corresponding to four different substance-use prevention programs. One group received information on the consequences of use (ICU), which can be regarded as a baseline or control program. A second group received this baseline program plus resistance training (ICU + RT), designed to build skills to help students resist future offers of substance use by their peers. A third group received the baseline plus normative education (ICU + NORM), designed to correct students' perceptions about the prevalence and acceptability of use. The fourth group received all three program components (ICU + RT + NORM).

The goal of this analysis will be to examine the possible effects of the grade 7 intervention on reported alcohol use in grade 9, controlling for important covariates measured in grade 7. This exercise is designed to illustrate the usefulness of our software and the MI paradigm, and to raise certain points about imputation modeling. To keep matters simple, we will sidestep two issues that complicate these data. The first issue pertains to the distribution of reported alcohol use. In both seventh and ninth grades, a substantial proportion of students reported no alcohol use at all, whereas the remaining students reported levels ranging from very little use to heavy use. Consequently, the distribution of this variable is "semicontinuous," a mixture of zeros and continuously varying positive responses. Ideally, the statistical analysis and missing-data procedure should address this variable's two-part nature, but doing so in the present article would be an unwelcome digression. For now, we will simply collapse alcohol use into a dichotomous variable (some use versus no use), sacrificing our ability to distinguish among users of varying degree.

The second issue that we will not address pertains to the fact that the subjects in this study may not act independently of one another. The students are grouped into schools,

and the intervention programs were administered at the school level. An ideal analysis of these data should account for their grouped or multilevel structure. Methods for multilevel regression analysis are readily available (e.g. Bryk & Raudenbush, 1992), and imputation procedures for multivariate multilevel data have been developed by Schafer (submitted). For simplicity, however, we will ignore the multilevel aspects of these data and use methods that treat the students as independent agents. Our failure to account for the multilevel structure means that the precision of estimated relationships will be somewhat overstated, and that the actual statistical significance of an effect is somewhat less than what we will report.

The variables included in the analysis and their rates of missingness are reported in Table 2. All variables except for ALC9 were collected in seventh grade prior to intervention. Notice that the variables PARRELAT, MONITOR, PEERPREV and GRADES7 are missing for approximately one-third of the subjects. Most of these missing values were planned, a consequence of Graham's three-form design. The missingness for ALC9, however, is due to a combination of attrition, absenteeism, and insufficient time to complete the questionnaire when the ninth grade measures were taken. Histograms of the nine variables are displayed in Figure 2.

One interesting feature of these data is that GRADES7 is a fairly strong predictor of missingness for ALC9. Students with lower grades tended to have poorer reading skills which hindered their ability to complete the questionnaire in time and, consequently, led to missing values for ninth grade alcohol use. To the extent that GRADES7 explains part of the missingness for ALC9, including it in the imputation procedure tends to make the MAR assumption more plausible. Variables thought to be strong predictors of missingness should be included in the imputation procedure wherever possible, even if they are not needed for later substantive analyses.

Table 2: Variables from the AAPT study, with rates of missing values

Name	Description	missing (%)
PARRELAT	relationship with parents (↑ means better relationship)	35
MONITOR	parental monitoring (↑ means more monitoring)	35
PEERPREV	prevalence of use among peers (↑ means more perceived use)	36
SMOKE7	smoking, grade 7 (↑ means more smoking)	0
ALC7	alcohol use, grade 7 (0 = none, 1 = some)	0
FRNDUSE	substance use by friends (↑ means more use)	0
GRADES7	grades, grade 7 (↑ means better reading skills)	37
SEX	gender (0 = female, 1 = male)	0
ALC9	alcohol use, grade 9 (0 = none, 1 = some)	45

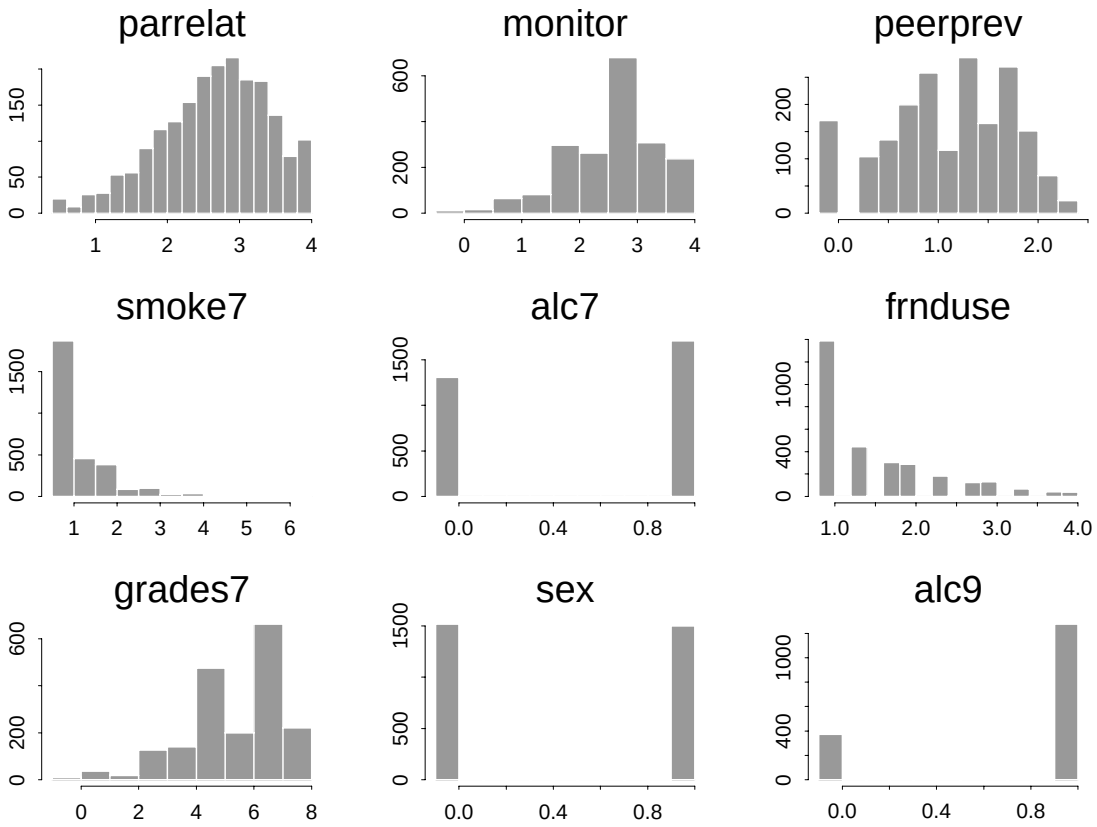


Figure 2: Histograms of nine variables from the AAPT study

By applying the MAR assumption, we are in effect assuming that any unmeasured covariates related to missingness are essentially unrelated to the key variables in this study. Because the missingness for ALC9 was uncontrolled, we have no way to support or discredit this hypothesis directly. Extensive follow-up evaluations, however, suggested that in the vast majority of cases, these data were missing for reasons unrelated to substance use (Graham, Hofer, & Piccinin, 1994). Therefore, we believe that applying the MAR assumption to these missing values is not entirely unreasonable.

5.2 Plan for analysis

Prior to imputation, it is helpful to have in mind the methods by which the data will ultimately be analyzed, to devise an imputation model that will be compatible with the intended analyses.

Because our main goal is to assess possible program effects on alcohol use at grade 9, we will contruct a logistic regression model to predict ALC9. Letting p denote the probability that a subject reports alcohol use (ALC9 = 1), we will fit a model of the form

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q,$$

where $X_1, \ldots, X_q$ represent the other covariates listed in Table 2 and dummy indicators to distinguish among the program groups. Because there are four program groups, let us define three indicators: RTICU (equal to one if the student received ICU + RT, zero otherwise); NORMICU (one if ICU + NORM, zero otherwise); and COMBICU (one if ICU + RT + NORM, zero otherwise). Under this coding scheme, the β-coefficients for these indicators will distinguish the rate of alcohol use in each of the three experimental treatment groups (ICU + RT, ICU + NORM, and ICU + RT + NORM, respectively) from that of the control

group (ICU only).

For simplicity, we will not investigate possible interactions among the program indicators and other covariates. That is, we will determine only whether the programs appeared to have any discernible effects on ALC9 in an overall sense, and we will not attempt to measure any differences in program effects among subgroups (e.g. boys and girls). Thus it will be sufficient to use an imputation model that allows simple associations or correlations between each dummy indicator (RTICU, NORMICU, and COMBICU) and ALC9. If we had intended to look at interactions of program and gender, we would need an imputation model that allows more general three-way associations between the dummy indicators, ALC9, and gender.

5.3 The imputation model

Because this dataset contains both continuous and binary variables, we might consider imputing the missing data using the program MIX. However, it is also possible to do an acceptably good job using the simpler package NORM, which treats the variables as if they are jointly normal. The binary variables ALC7, SEX, and ALC9 can be imputed under normality assumptions and then rounded off to zero or one, a procedure that tends to work quite well in practice Schafer (1997). The variables SMOKE7 and FRNDUSE are highly skewed and clearly nonnormal. One might consider transforming these variables to make the normality assumption more plausible. The rates of missingness for these variables are extremely low (less than 0.5%), however, so they will almost never be imputed; including them without transformation will produce very little distortion of their distributional shapes. They are present mainly to preserve their relationships with other variables, particularly ALC9. To

preserve program effects, the dummy indicators RTICU, NORMICU and COMBICU are also added to the model, even though they are completely observed and do not require imputation.

To understand the properties of a dataset imputed with NORM software, one must understand the essential features of the multivariate normal model. Unless special restrictions are added to the covariance structure (and NORM does not apply any such restrictions), the multivariate normal model implies that each variable has an additive linear regression on all other variables with main effects but no interactions. As a result, NORM imputes data using linear regression predictions at the subject level. For example, consider a subject with a missing value for PEERPREV, but observed values for all other variables. Her value for PEERPREV is imputed from a linear regression equation based on her recorded values for all other variables. The imputed value is not merely a regression prediction, but incorporates an appropriate level of residual noise to preserve the dataset's covariance structure. If two or more variables are missing for a subject, then they are imputed under a multivariate regression model that includes residual correlations among them. Imputations from NORM are thus created under a system of simultaneous regression models in which each variable potentially depends on all other variables.

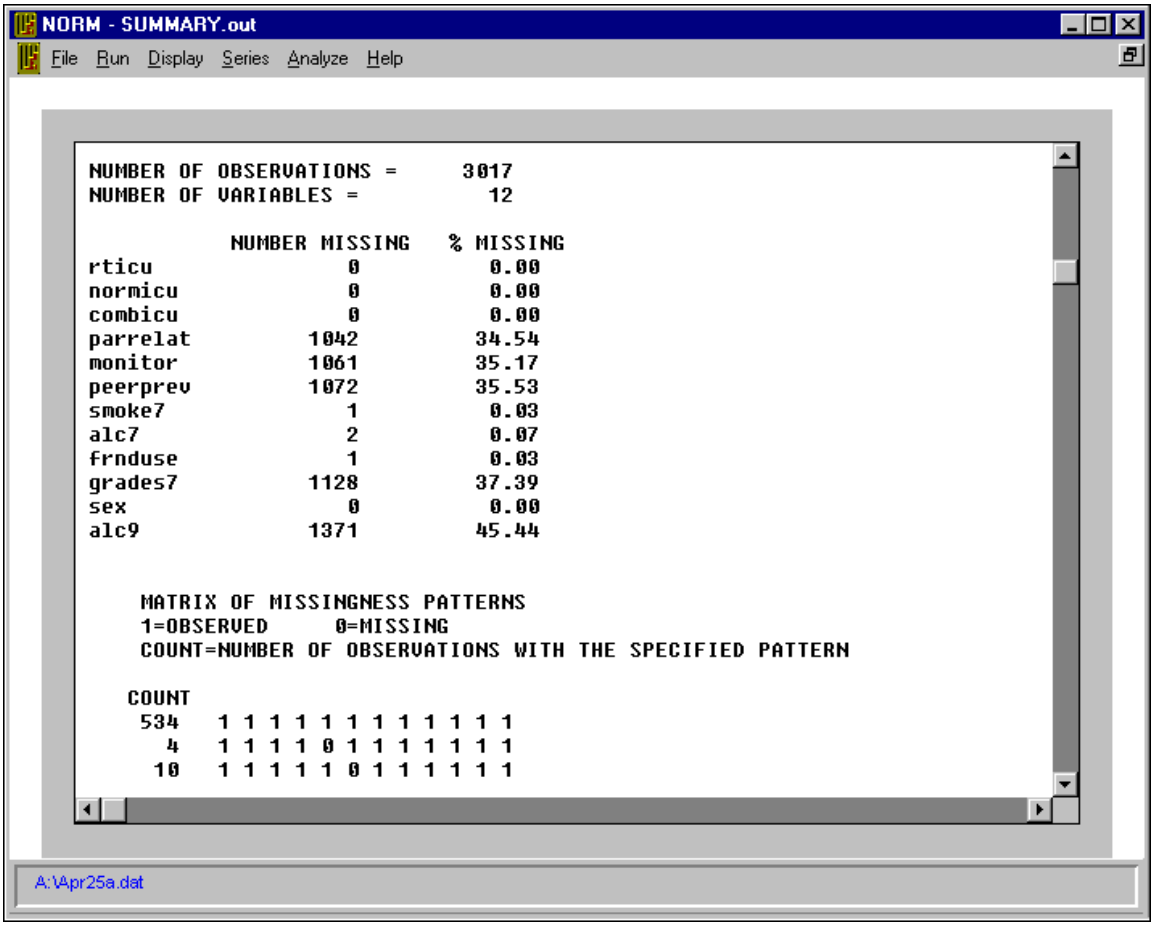
Notice that the logistic regression that will ultimately be used for analysis is nonlinear and therefore does not precisely mesh with the linear model applied by NORM. Simulation work (e.g. Schafer, 1997, ch. 6) has shown that the biases that may arise due to discrepancies such as these—slight differences in functional form between the imputer's and analyst's models—are usually minor and have little or no discernible impact on the statistical analysis. A more serious inconsistency would arise, however, if the subsequent regression model for ALC9 included interactions among the predictors. The multivariate normal model allows

simple pairwise correlations among variables, but more complex associations such as interactions are not supported. Straightforward imputation under a normal model would tend to dampen interactions, making them more difficult to detect in post-imputation analyses.

If interactions are to be a crucial part of post-imputation analyses, then one should apply an imputation model that preserves the interactions of interest. Higher-order associations among variables can be specified in the models used by CAT and MIX. It is also possible to preserve higher-order associations within NORM by including products of variables. For example, suppose that we wanted to investigate differences in program effects among boys and girls. Adding the product variables RTICU $\times$ SEX, NORMICU $\times$ SEX, and COMBICU $\times$ SEX to the NORM imputation procedure would ensure that program by gender interactions on ALC9 would be preserved.

5.4 Summarizing the data

The first step in using NORM is to prepare the data set. NORM reads ordinary text (ASCII) files where each line represents data for a different subject and the variables are separated by one or more blank spaces. NORM prompts the user for the number of variables, the number of cases or subjects, and a numerical missing value code. For obvious reasons, missing values should be denoted by a number which is not a plausible response for any of the variables (e.g. -999). Once the data have been read, NORM may be asked to display a printed summary of the observed data. This summary includes the number and percent missing for each variable, as well as the means and standard deviations of the observed data. It also provides a matrix of 0's and 1's which indicate missingness patterns in the data. A portion of the NORM window which summarizes the AAPT data is shown in Figure 3.



```
NORM - SUMMARY.out
File  Run  Display  Series  Analyze  Help

NUMBER OF OBSERVATIONS =     3017
NUMBER OF VARIABLES =          12

           NUMBER MISSING   % MISSING
rticu                0          0.00
normicu              0          0.00
combicu              0          0.00
parrelat          1042         34.54
monitor           1061         35.17
peerprev          1072         35.53
smoke7               1          0.03
alc7                 2          0.07
frnduse              1          0.03
grades7           1128         37.39
sex                  0          0.00
alc9              1371         45.44


    MATRIX OF MISSINGNESS PATTERNS
    1=OBSERVED     0=MISSING
    COUNT=NUMBER OF OBSERVATIONS WITH THE SPECIFIED PATTERN

   COUNT
    534   1 1 1 1 1 1 1 1 1 1 1 1
      4   1 1 1 1 0 1 1 1 1 1 1 1
     10   1 1 1 1 1 0 1 1 1 1 1 1

A:\Apr25a.dat
```

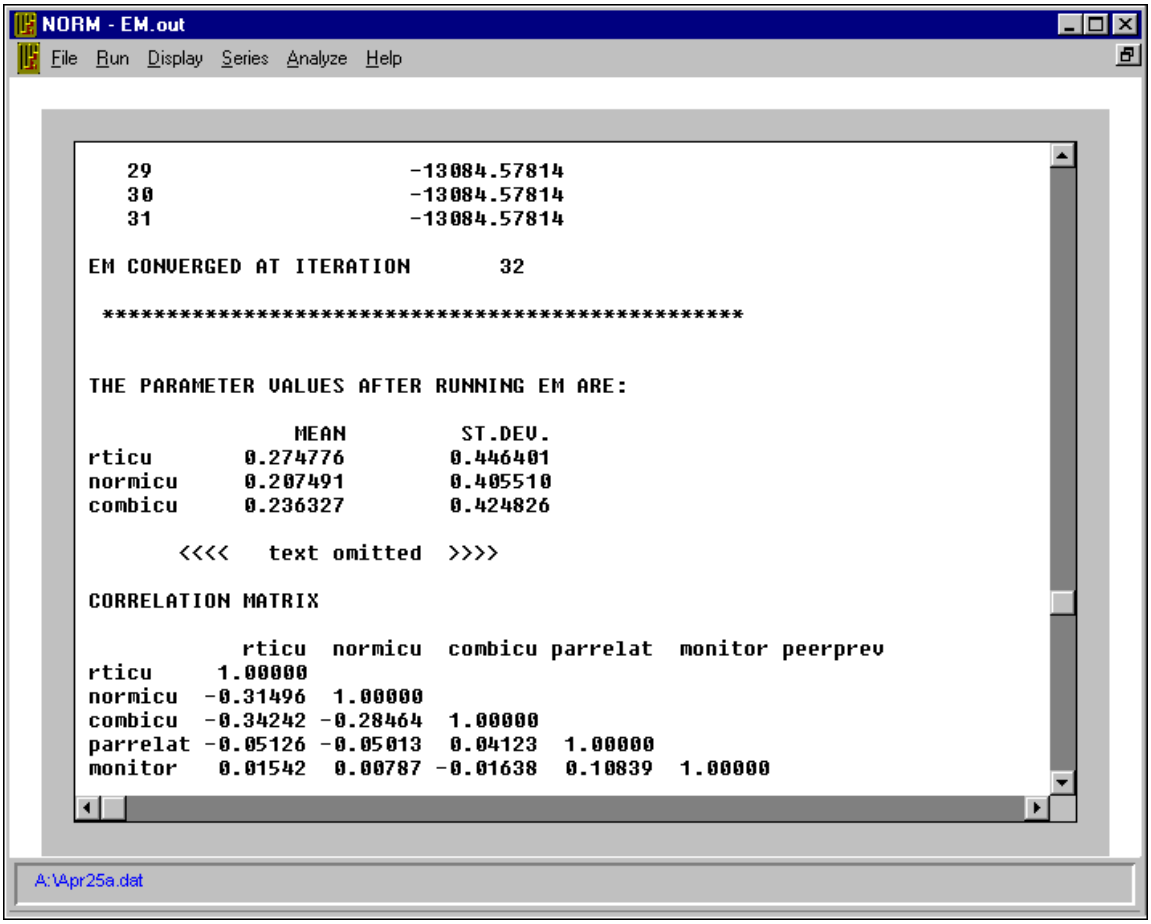
Figure 3: NORM data summary window showing rates of missingness, patterns of missingness, and means and standard deviations by variable

It is interesting to note that only 534 of the 3,017 subjects have complete data on all variables. If an analysis involving all these variables were performed using standard statistical packages, the built-in case deletion procedures would discard up to 82% of the subjects, resulting in a substantial loss of power.

5.5 Running EM and DA

After examining the data summary, we ran the EM algorithm to compute maximum-likelihood estimates of parameters (means and covariances) under the normal model. The procedure converged in 32 iterations, which took less than 10 seconds on a 133 MHZ Pentium computer. Upon convergence, NORM provides an iteration history, including the observed-data loglikelihood, as well as the final estimates of the means, standard deviations and correlations, as shown in Figure 4. The estimated parameters are automatically saved to a file in an appropriate format to serve as starting values for subsequent runs of data augmentation.

Based on the rapid convergence of EM, it appeared that 32 cycles of data augmentation would be sufficient for DA to converge in distribution. For an extra margin of safety, we decided to double that number and take 64 cycles of DA between imputations. We ran DA for a total of 640 cycles, producing an imputation at every 64th cycle for a total of $m = 10$ imputations. We used the default noninformative prior distribution and also stored the parameters from each cycle of DA so that we could later create time-series and ACF plots. The entire data augmentation and imputation procedure took 127 seconds. Each imputed dataset is stored as a text (ASCII) file, in a format similar to that of the input data.



```
NORM - EM.out
File  Run  Display  Series  Analyze  Help

   29                    -13084.57814
   30                    -13084.57814
   31                    -13084.57814

EM CONVERGED AT ITERATION       32

 ***************************************************


THE PARAMETER VALUES AFTER RUNNING EM ARE:

                MEAN          ST.DEV.
rticu       0.274776        0.446401
normicu     0.207491        0.405510
combicu     0.236327        0.424826

       <<<<   text omitted  >>>>

CORRELATION MATRIX

            rticu  normicu  combicu parrelat  monitor peerprev
rticu     1.00000
normicu  -0.31496  1.00000
combicu  -0.34242 -0.28464  1.00000
parrelat -0.05126 -0.05013  0.04123  1.00000
monitor   0.01542  0.00787 -0.01638  0.10839  1.00000

A:\Apr25a.dat
```

Figure 4: NORM window showing output from a run of the EM algorithm

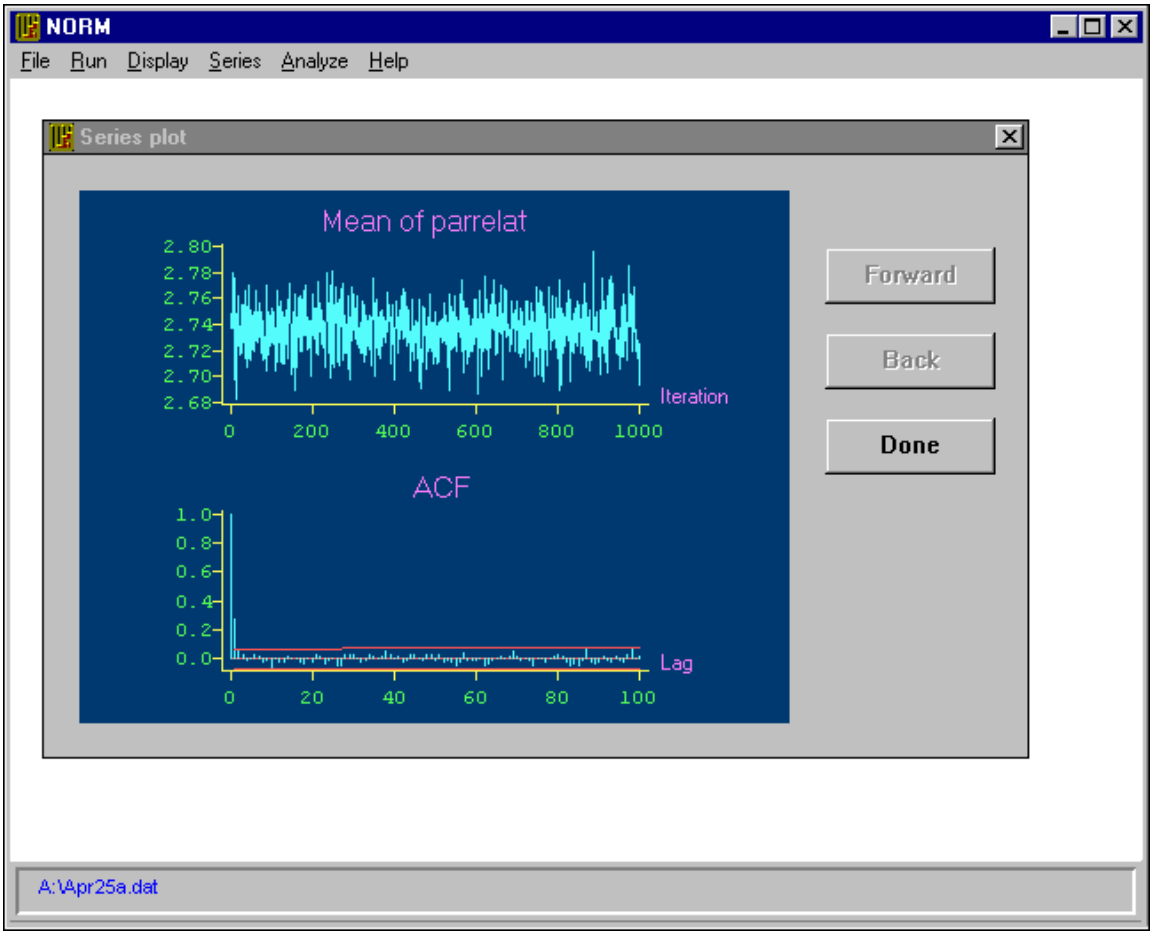


Figure 5: Time-series and ACF plot from a data augmentation run showing rapid convergence

5.6 Assessing convergence

As discussed earlier, it is a good idea to examine time-series and ACF plots of the parameters to diagnose convergence. Plots for one of the parameters are shown in Figure 5. In the time series, the value of the parameter (vertical axis) is plotted against the iteration number (horizontal axis). If DA converges quickly, the series should resemble a horizontal band without long upward or downward trends, as it does have. Long-term drifts or trends indicate positive lag-k correlations at large values of k. These correlations, as displayed in the ACF plot, die out very quickly for this parameter. Plots for all other parameters show

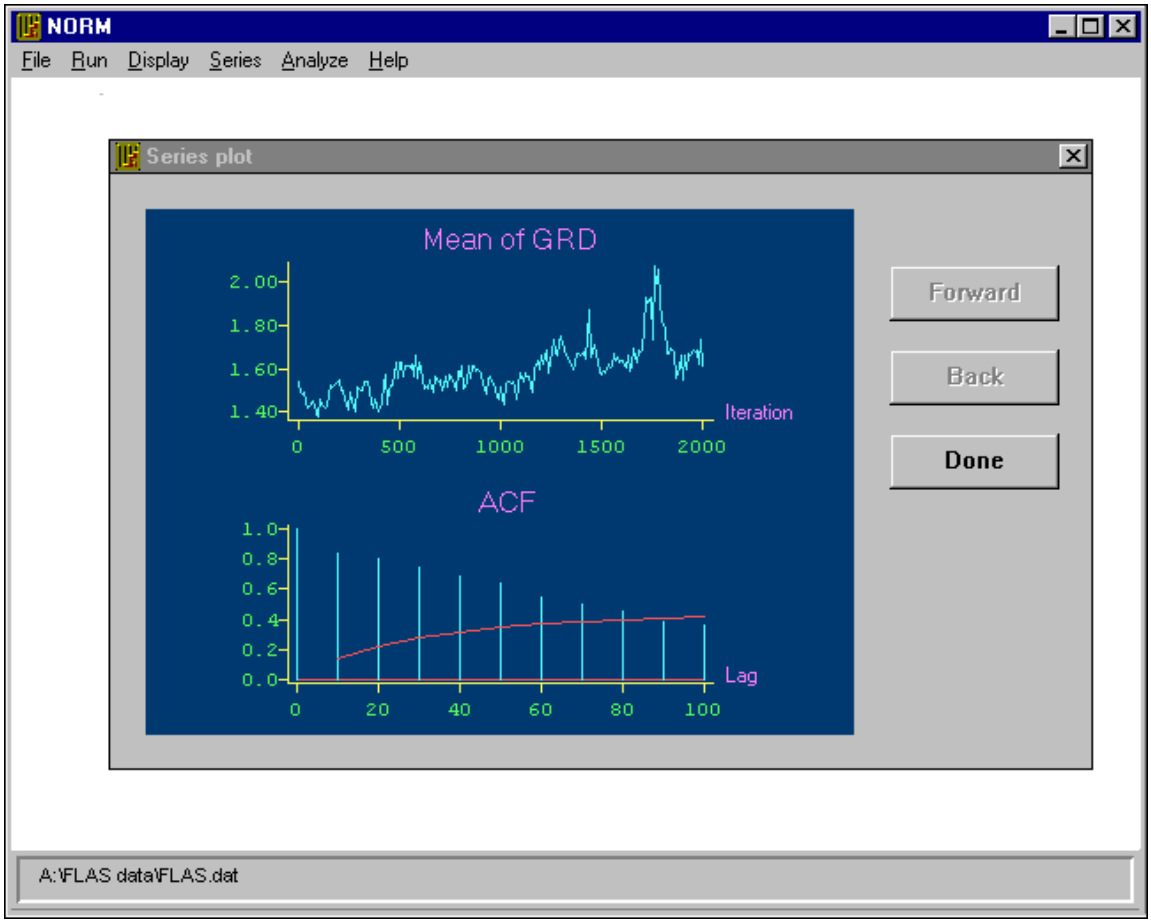


Figure 6: Time-series and ACF plot from a data augmentation run showing pathological behavior

similar behavior, verifying that the DA algorithm converged rapidly in this example.

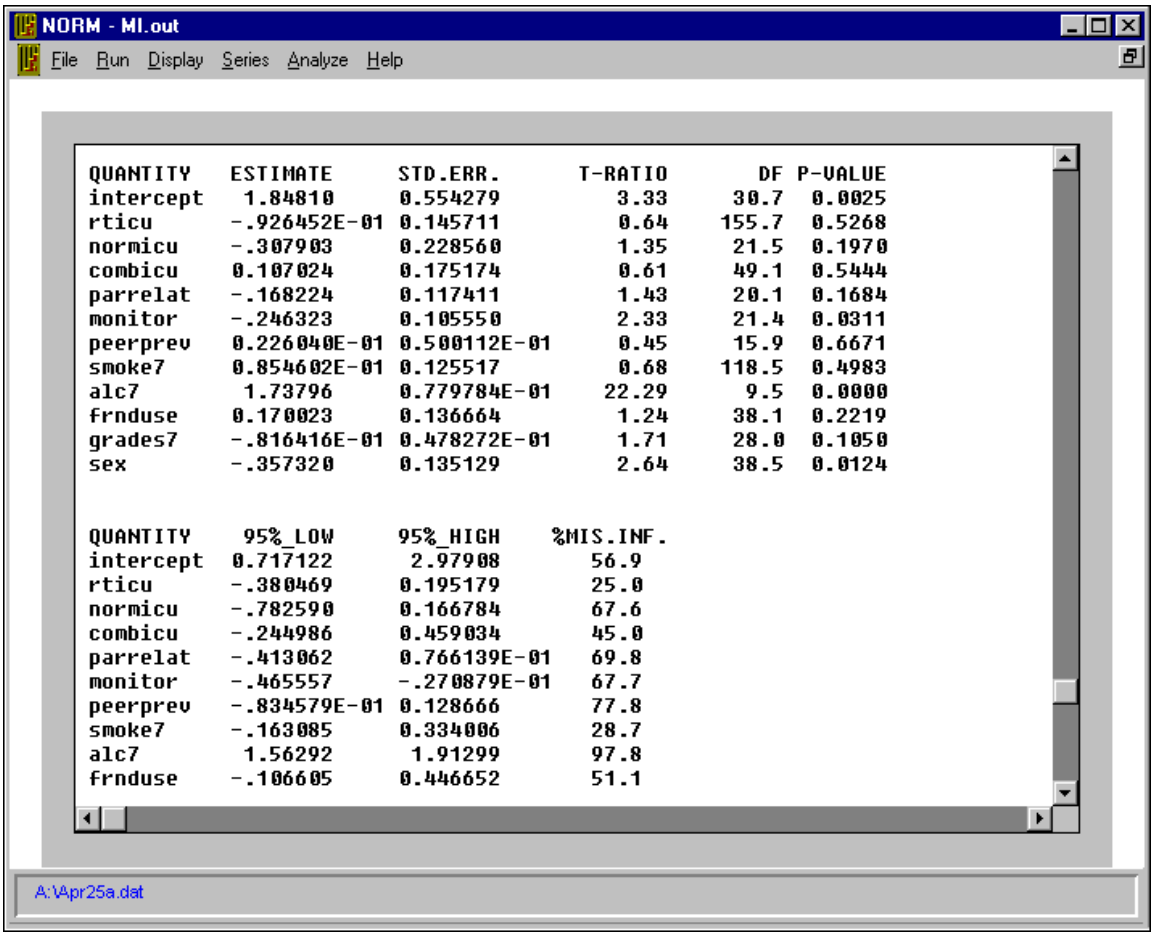
For comparison, plots for a parameter from a poorly behaved DA run are shown in Figure 6. The data and model that produced this plot are discussed by (Schafer, 1997, Sec. 6.3) and will not be explained here. For our purposes, we simply note that the long-term drift and high autocorrelations imply a pathological situation with extremely high rates of missing information. In fact, the rate of missing information for this parameter is 100%. For further explanation and remedies for such unusual situations, see Chapters 3, 4 and 6 of Schafer (1997).

5.7 Analyzing the imputed datasets

Following imputation, we performed logistic regression analyses to predict ALC9. The analyses were carried out in SAS using PROC LOGISTIC. Our SAS program read in each imputed dataset and rounded off the imputed values of ALC7, SEX, and ALC9 to zero or one. A logistic regression was fit to predict ALC9 from all other variables, including the dummy indicators RTICU, NORMICU, and COMBICU. The estimated coefficients and standard errors were then saved to a text file. This text file consists of a column of coefficients and a column of standard errors separated by blank space. The same SAS program was run ten times, with only trivial modifications to change the input file names to `AAPT*.IMP` and the output file names to `RESULT*.DAT` for $* = 1, 2, \ldots, 10$.

5.8 Combining the results

The final step in our analysis was to combine the coefficients and standard errors in the `RESULT*.DAT` files according to Rubin's rules. The NORM program carries out this step automatically; after restarting NORM, select "MI inference" under the "Analyze" menu option. The printed report generated by NORM, as shown in Figure 7, resembles a table of coefficients produced by traditional regression software. For each coefficient, the report provides the overall estimate $\bar{Q}$, the standard error $\sqrt{T}$, the degrees of freedom df for the t-approximation, and the p-value for testing the hypothesis $Q = 0$ against the two-sided alternative. A second table provides the lower and upper limits of a 95% confidence interval and the estimated percent missing information for each coefficient. The high rates of missing information for all coefficients reflect the high percentage of missing values for the response variable ALC9.



NORM - MI.out

File Run Display Series Analyze Help

QUANTITY	ESTIMATE	STD.ERR.	T-RATIO	DF	P-VALUE
intercept	1.84810	0.554279	3.33	30.7	0.0025
rticu	-.926452E-01	0.145711	0.64	155.7	0.5268
normicu	-.307903	0.228560	1.35	21.5	0.1970
combicu	0.107024	0.175174	0.61	49.1	0.5444
parrelat	-.168224	0.117411	1.43	20.1	0.1684
monitor	-.246323	0.105550	2.33	21.4	0.0311
peerprev	0.226040E-01	0.500112E-01	0.45	15.9	0.6671
smoke7	0.854602E-01	0.125517	0.68	118.5	0.4983
alc7	1.73796	0.779784E-01	22.29	9.5	0.0000
frnduse	0.170023	0.136664	1.24	38.1	0.2219
grades7	-.816416E-01	0.478272E-01	1.71	28.0	0.1050
sex	-.357320	0.135129	2.64	38.5	0.0124

QUANTITY	95%_LOW	95%_HIGH	%MIS.INF.
intercept	0.717122	2.97908	56.9
rticu	-.380469	0.195179	25.0
normicu	-.782590	0.166784	67.6
combicu	-.244986	0.459034	45.0
parrelat	-.413062	0.766139E-01	69.8
monitor	-.465557	-.270879E-01	67.7
peerprev	-.834579E-01	0.128666	77.8
smoke7	-.163085	0.334006	28.7
alc7	1.56292	1.91299	97.8
frnduse	-.106605	0.446652	51.1

A:\Apr25a.dat

Figure 7: NORM window displaying combined results from ten logistic regression analyses

Because this analysis does not reflect the multilevel structure of the data, the p-values shown in Figure 7 are probably "too significant," overstating the strength of evidence for each coefficient. Nevertheless, the estimated coefficients for MONITOR, ALC7, and SEX are highly significant and the effects are in the expected direction. Alcohol use at grade 9 appears to be negatively associated with parental monitoring and positively associated with alcohol use at grade 7, and girls are less likely to use alcohol than boys. None of the dummy indicators for treatment are significantly different from zero. The estimated coefficient for NORMICU is negative and fairly large, however, suggesting that the ICU + NORM intervention may have some benefit in reducing alcohol use relative to the baseline ICU program. This finding is rather consistent with previous analyses of AAPT by Hansen & Graham (1991) and Palmer et al. (submitted), who reported that the only group that behaved significantly better than ICU was NORM + ICU. The fact that the NORMICU effect seems weaker in this present analysis may be due to the information lost by our dichotomization of ALC9. Grouping the reported users of alcohol into a single category eliminates the possibility of detecting program effects that diminish the levels of use among users.

6 Discussion

The analysis of datasets with missing values is one area of statistical science where real advances have recently been made. Modern missing-data techniques which substantially improve upon old ad hoc methods are finally becoming available to data analysts. Among these new techniques, multiple imputation is especially powerful because of its generality. Standard programs for data analysis such as SAS, SPSS, and LISREL were never intended

to handle datasets with a high percentage of incomplete cases, and the missing-data procedures built into these programs are crude at best. On the other hand, these programs are exceptionally powerful tools for *complete* data, providing an amazing variety of exploratory methods and model-fitting routines. Our new software for MI supplements rather than replaces these statistical packages. MI does resemble the older methods of case deletion and ad hoc imputation in that it addresses the missing-data issue at the beginning, prior to the substantive analyses. Unlike the ad hoc methods, however, MI solves the missing-data problem in a principled and statistically defensible manner, incorporating missing-data uncertainty into all summary statistics.

MI is not the only modern missing-data tool to become available to researchers. Some producers of statistical software are beginning to incorporate incomplete-data features directly into certain types of model-fitting routines. These procedures are similar to MI in that they implicitly average over a predictive distribution for the missing values, but the averaging is performed using analytic or numerical methods rather than simulation. Programs for multilevel regression modeling, including HLM (Bryk, Raudenbush, & Congdon, 1996) and SAS PROC MIXED (Littell et al. 1996) allow arbitrary patterns of missing values in the response variable. Two programs for structural equations modeling, Mx (Neale, 1991) and Amos (Arbuckle, 1995), can perform direct maximum-likelihood estimation using both complete and incomplete cases. With reasonably large sample sizes, these direct maximum-likelihood methods should lead to essentially the same results as MI. In fact, direct maximum-likelihood methods will be slightly more efficient than MI because they do not rely on simulation.

To the extent that direct maximum-likelihood methods are available, we wholeheartedly encourage analysts to use them. For many types of analyses, however, no direct pro-

cedures currently exist. For example, we know of no statistical software capable of fitting a logistic regression model with missing values on both the predictors and the response, as in the AAPT example above. It appears unlikely that any such software will become available in the forseeable future. Direct maximum-likelihood methods are computationally complicated and require a special implementation for each new type of model. MI, on the other hand, is a general technique that can be applied to a wide variety of modeling problems right now. We look forward to developing MI routines and software for wider classes of incomplete-data problems, and to disseminating these methods beyond a small circle of statistical experts to the wider community of researchers and data analysts.

7 References

Arbuckle, J.L. (1995). *Amos Users' Guide*. Small Waters, Chicago.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Sage, Newbury Park.

Bryk, A.S., Raudenbush, S.W., and Congdon, R.T. (1996). *Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs*. Scientific Software International, Chicago.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A., Rubin, D.B., & Schafer, J.L. (1995). A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. In *Proceedings of the Annual Research Conference*, pp. 257–266. Bureau of the Census, Washington, D.C.

Gilks, W.R., Richardson, S., & Spiegelhalter, D.J. (Eds.). (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Graham, J.W., Hofer, S.M., & Piccinin, A.M. (1994). Analysis with missing data in drug prevention research. In Collins, L.& Seitz, L. (Eds.), *National Institute on Drug Abuse Research Monograph Series*, Vol. 142, pp. 13–62. National Institute on Drug Abuse, Washington, D.C.

Graham, J.W. & Schafer, J.L. (in press). On the performance of multiple imputation for multivariate data with small sample size. In Hoyle, R. (Ed.), *Statistical Strategies for Small Sample Research*. Sage, Thousand Oaks.

Hansen, W.B. & Graham, J.W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Preventive Medicine*, *20*, 414–430.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996). *SAS System for Mixed Models*. SAS Institute, Cary, NC.

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

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

Neale, M.C. (1991). *Mx: Statistical Modeling.* Available from M.C. Neale, Box 3, Department of Human Genetics, Medical College of Virginia, Richmond, VA.

Palmer, R.F., Graham, J.W., Taylor, B., & Tatterson, J.W. (submitted). Interpreting latent variable measurement models in health behavior research. *Journal of Behavioral Medicine*.

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

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

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

Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the*

American Statistical Association, *91*, 473–489.

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

Schafer, J.L. (submitted). Imputation of missing covariates under a multivariate linear mixed model. *Biometrics*.

Schafer, J.L., Khare, M., & Ezzati-Rice, T.M. (1993). Multiple imputation of missing data in NHANES III. In *Proceedings of the Annual Research Conference*, pp. 459–487. Bureau of the Census, Washington, D.C.

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

Author Note

Joseph L. Schafer, Assistant Professor; Maren K. Olsen, Graduate Assistant; Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802. This research was supported by grant 1-P50-DA10075-01, National Institute on Drug Abuse. Special thanks to John Graham for providing data from the Adolescent Alcohol Prevention Trial and advice on their analysis.