

Last revised: 1/24/03

Draft: Please do not cite, quote or distribute without permission

TO: Jay Bell
FROM: Susan Ettner
CC: Cheryl McDonnell, Carolyn Lichtenstein, Steve Banks, Alfonso Ang, Chris Conover, Marcia Weaver
RE: Missing Data Handling

The purpose of this memo is to bring the attention of the multi-site investigators to the issue of missing data, in the hope of reaching some consensus on the approach taken. The memo will describe the missing data problem, lay out the pros and cons of potential approaches to this problem, and offer recommendations.

Missing data is a common but serious problem in many health care studies. The problems are manifested as a loss of efficiency (i.e., the parameter of interest is estimated with less precision) or even possible bias (i.e., the magnitude or sign of the parameter estimate are incorrect). To best utilize all of the available information and get unbiased study results, we need to understand the missing data mechanism and implement the appropriate method to handle the missing data. In determining which missing data mechanism applies, the most important distinction is whether the data appear to be missing at random or follow some pattern. All of the commonly used methods for dealing with missing data require, at a minimum, the assumption that the data are “missing at random (MAR),” that is, the probability of missing data for a particular variable does not depend on the value of the variable after controlling for the other variables in the model. (See the Technical Appendix for further detail on this issue and examples.) An exposition of the techniques for dealing with missing data in the absence of this assumption is outside the scope of the current memo, since these methods become complicated very quickly.

The main options for dealing with missing data are case deletion and imputation, including single imputation methods (e.g., mean imputation, hot-decking, cold-decking) and multiple imputation (Little and Rubin, 1987). Besides the nature of the missing data, the most suitable method will depend on sample size, the researcher’s familiarity with different imputation methods, availability of statistical software, and ease of interpretation of the final results.

The most straightforward method for handling missing data is case deletion. Many statistical software packages automatically omit from analysis all cases that have a missing value for any variable (either dependent or independent). With case deletion, no additional provision for missing data is made in the subsequent analysis after the data set has been altered. The research usually proceeds as if the omitted cases had never really been observed.

The desirability of the case deletion method depends on the extent to which the incompletely observed cases differ systematically from the completely observed ones, as well as the amount of missing data. If the incomplete cases (those being deleted) are different from the complete cases in important ways that cannot be controlled in the regression model, then their omission will bias the results. For example, suppose we are interested in examining treatment adherence, and that adherence depends in part on the patient’s level of “defiance,” an attribute that cannot be measured in our dataset. If defiant patients are simultaneously less likely to be adherent to their treatment regimens and less likely to be compliant in answering survey questions, then excluding those patients from the sample may result in misleading estimates of treatment adherence. Although this problem may be less obvious when the goal is to estimate a relationship between two variables (e.g., treatment adherence and health status), bias may still result if we cannot fully adjust for the fact that the deleted cases were different.

Even if the cases with missing data values are a completely random subset of the entire sample, deletion will result in a smaller sample and hence reduce power. Because cases with missing values for *any* variable used in the analysis need to be deleted, the reduction in sample size can often be substantial. For these reasons, case deletion is not a recommended strategy for studies in which small sample size and low power are concerns.

An alternative approach is to impute some reasonable data to replace the missing data, and then implement standard statistical analyses. Imputation is popular because it is conceptually simple and because the resulting sample has the same number of observations as the full data set. If the imputation method is “proper,” these analyses will yield “correct” results, i.e., the estimated regression coefficients and standard errors will be close to the true values in large samples.

With single imputation methods, one essentially fills in the missing values. The data used to fill in the missing values can come from a variety of sources, depending on the method used. Similarly to the case deletion method, a strong advantage of single-imputation methods is that they result in a single database (in this case, with the full sample and imputed values replacing the missing values) and the new database is treated as though the imputed values were real. Thus all of the investigators on a study work with the same database, enhancing consistency and comparability of the results across analyses. The disadvantage is that single-imputation methods can lead to biased estimates, either of the regression coefficient or of the variance.

For example, with conditional mean imputation, one would replace the missing value for a particular variable with a predicted value from a regression of that variable on the other covariates, using the subsample of observations with complete data. As a result of plugging in a regression prediction for each person with missing data, the new, imputed values will demonstrate less variability than real data values would have. If the regression of interest is then estimated using these data, the standard errors will be underestimated, and the statistical significance of the regressor effects overstated. With low rates of missing data, the bias may not be large, but when data must be imputed for a substantial number of observations, the bias will be noticeable.

Random error can be added to the imputed values, for example, by randomly drawing an error term from a normal distribution and adding it to the imputed value. However, the standard errors from the estimation model will still be too low, as they will not account for the fact that different random draws of the error term would have produced different results. They also will not account for variability resulting from the fact that the parameters of the imputation model are only estimates themselves and therefore measured with uncertainty as well.

The idea behind multiple imputation is essentially to perform the imputation procedure a number of times (typically anywhere from 3 to 10 times) in order to estimate the variability across imputations and take it into account in determining the variability around the final estimates. Depending on how the imputation is performed, additional uncertainty resulting from the estimation of the parameters in the imputation model can be incorporated as well.

In contrast to single imputation, with multiple imputation, there is not a single data value assigned to replace the missing data and therefore there is not a single imputed database. Instead, one creates multiple datasets (say five), each with slightly different imputed data. The model of interest is estimated five times, once for each of the datasets, and then the results of all of those runs (e.g., the estimates and standard errors) are combined in order to get a final averaged estimate of the parameter of interest and the variance around this estimate. The technical appendix describes the rules used to combine the estimates and calculate the variance.

For example, if one wanted to look at the impact of income on service use, and the income variable has missing data, then the service use regression would be estimated five times, each time based on a slightly different imputed value for income. The final estimated coefficient for income would be the average of the five coefficient estimates. The variance of this estimate would take into account both the average of the variances estimated in each of the five runs, plus the variance of the coefficient estimates across the five runs. If the five different regressions yielded almost identical estimates for the income coefficient, then the final estimate of the variance for the income coefficient would essentially just be the average variance from the five regressions. On the other hand, if the estimated income coefficient varies dramatically across the five regressions, then the variance estimate is increased to take into account the uncertainty resulting from the imputation.

Although multiple imputation is widely considered to be better than single-imputation from a theoretical point of view, it does have disadvantages, including the computing resources and potential complexity involved with analyzing multiple databases and combining the estimates. Although the software is not yet optimal for handling all situations, standard software packages are beginning to include modules for performing multiple imputation. SAS has a “beta” version of a procedure that will create the multiply imputed databases, run regression models with the databases and combine the results into a single estimate. However, the canned command is limited in terms of the types of regression models that can be run, so additional programming may be necessary.

Other problems with multiple imputation are that the missing data values are imputed simultaneously, so it would be difficult to impute more than a handful of variables, given the sample size. It is also more complicated – though not impossible -- to impute variables that are not continuous. Despite these limitations, multiple imputation is strongly preferred by statisticians and it is considered to be state of the art from a scientific point of view.

Additional issues

One choice that needs to be made is when to impute missing values for a variable, and when to simply delete observations if data for the variable are missing. Normally imputation is done for the independent variables, using the information contained in both the other independent variables and the dependent variables. It is less common to impute the missing values in dependent variables, although statisticians advocate this approach in certain cases (see Technical Appendix for details). Some study teams choose to impute values only for the control variables but not the main exposure variables. Another option is to treat imputation as a sensitivity analysis, comparing the results of models estimated using imputation for the main exposures to models estimated after dropping the observations for which the data for the main exposures are missing.

Another consideration is that our study is longitudinal, so strategies should be chosen in order to handle the missing data in a comparable way for the baseline and followup surveys. Conventional MI methods treat observations as though they are independent from one another, so the imputation model must build in dependence between observations if we wish for the correlation to be preserved with the imputed values. An advantage of the longitudinal structure of the data is that it may be possible to use data from one wave to help impute missing data for other waves.

Multi-level models are an alternative way to deal with the issue of missing followup data when the entire followup survey is missing. However, this works only when the endpoint being examined is a “point in time” measure such as health status. For measures that are cumulative over time, such as services utilization or costs, one would still need to formally impute values for the missing waves and then add up the imputed and actual values for all of the waves over which the measure is being calculated.

Recommendations:

- 1) The original database, with all of the missing values, should be retained in case investigators wish to use their own preferred methods to deal with missing data for some or all of the variables. This allows investigators the flexibility to deal with missing data in the way that best fits their own particular analysis.
- 2) The Coordinating Center should distribute five datasets with multiply imputed values replacing the missing data for a set of variables agreed upon by the Steering Committee. By having access to datasets with complete data that are ready to analyze, investigators can gain the benefit of sophisticated multi-site approaches to missing data, even if they do not have the time, resources or desire to carry out complex imputation procedures themselves.
- 3) Investigators would have the option to perform five separate analyses with each of these databases and then combine the results into a single estimate, as described in the technical appendix. Alternatively, the investigators could perform an analysis using only one of the imputed datasets, chosen at random. To ensure consistency of results across research papers, the choice of which of the five datasets to use would be made ahead of time by the multi-site investigators, so that all researchers who wish to pursue the single-dataset approach would be using the same dataset. Note that both methodologies – analysis of a single dataset or analysis of all five -- require the assumption that there is no systematic bias, i.e., that the data are missing at random.
- 4) If some exploratory analysis is anticipated prior to the specification of the final models, the Coordinating Center may wish to create an additional, sixth dataset with imputed data. This dataset would be designated for doing any exploratory analyses that may be necessary; the other five datasets would be used for estimating the final model.

References:

Allison PD (2002). *Missing Data*. Sage University Papers. Series: Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage Publications, Inc.

Little RJA, Rubin DB (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schafer JL (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer JL, Olsen MK (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research* 33(4): 545-571.

Answers to frequently asked questions about multiple imputation are also available on Schafer's website: <http://www.stat.psu.edu/~jls/mifaq.html>

TECHNICAL APPENDIX

Missing data mechanisms

To understand the assumptions underlying different methods for dealing with missing data, it is necessary to gain familiarity first with the different possible mechanisms driving missing data. There are two general missing-data mechanisms: ignorable and non-ignorable (Rubin, 1976; Little and Rubin, 1987; Schafer, 1997). Missing data are said to be ignorable if the data are either *missing completely at random* (MCAR) or *missing at random* (MAR).¹

Suppose that we are trying to estimate the relationship between a variable with missing data, Y , and another variable, X . Missing data for Y are said to be *missing completely at random* if the probability that data on Y are missing does not depend on the value of either Y or X . In this case, the observations with complete data form a simple random subsample of the entire sample. Note that the probability that data on Y are missing can be related to the probability that data on X are missing, but it cannot be related to the actual *value* of X .

As an example, suppose that we are interested in estimating the association between CD-4 count and medications use. Suppose further that both of these variables are measured through chart review for patients chosen randomly from the original sample. Thus persons with missing data on medications are the same as those who are missing data on CD-4 count, yet they are still a random subsample and hence the data are MCAR. In contrast, suppose instead that data on medications use was collected directly from patients, and only those well enough to respond provided information. In that case, missing data on medications use is likely to be associated with worse CD-4 count, so the data cannot be considered MCAR.

Missing at random is a less stringent criterion, requiring only that the probability of missing data on Y cannot depend on the value of Y , after controlling for X . In other words, under MAR, missing values behave like a random sample of all values within subclasses defined by observed data. Y can be systematically higher or lower for non-respondents than for respondents, but after adjusting for X , the relationship between X and Y must be the same.

As an example, suppose that we are interested in estimating the association between income and minority group membership. If minority group members are less likely to report their income than non-minority group members, the data are no longer MCAR, but they may still be MAR. MAR would require that within the subsample of minority group members, those who report their income and those who do not have similar incomes on average. The same condition would have to hold within the subsample of non-minority group members.

An alternative way to think about these mechanisms in a regression context is that under MCAR, the probability of missing data on Y cannot depend on the values of either observable variables (X) or unobservable variables affecting Y (the error term of the regression). Under MAR, the probability of missing data on Y still cannot depend on the value of unobservable variables, although it is now allowed to depend on the value of observed variables. One implication is that MAR will be more plausible when all variables that predict missing data are included in the model, since excluding a variable essentially turns it into part of the error term.

¹ Another assumption, the distinctness of parameters, must also hold in order for the data to be ignorable. Distinctness of parameters means that the parameters for the model to be estimated (e.g., the regression model relating Y to X) are unrelated to the parameters for the missing data mechanism, so that the two processes can be modeled separately. Due to the implausibility that this condition would fail when MAR holds, most discussions of missing data mechanisms treat ignorability and MAR as being synonymous.

Unfortunately, because MCAR and MAR both depend on assuming something about unobservable variables, their validity cannot be tested. The majority of missing data methods used in empirical research simply assume ignorability and for many analyses, the assumption that the missing data mechanism is ignorable is a reasonable starting point. Although it is theoretically possible to model missing data under the assumption of non-ignorability (e.g., using sample selection models), it is neither straightforward nor common to use these methods in the health services research literature, so they are excluded from the current discussion.

Conventional methods for dealing with missing data

Methods for dealing with missing data can be classified into three categories: conventional methods (listwise deletion, pairwise deletion, dummy variable adjustment, marginal mean imputation, conditional mean imputation, hot-decking and cold-decking), maximum likelihood estimation and multiple imputation. The latter have become more popular in recent years, as software options have improved. The following is a brief summary of some of the conventional methods that have been commonly used in the past.

List-wise deletion (complete-case analysis, case deletion) List-wise deletion involves deleting from the sample all observations with missing data for any of the analysis variables. Some of the advantages of this approach are that it can be used for any type of statistical analysis and requires no special methods. If the missing data are MCAR, then standard statistical analysis can be used with the complete-case data without worrying about bias, either of the coefficient estimates or the variance. In this case, the only disadvantage is the loss in power due to the decrease in sample size. However, if the proportion of the sample with any missing data is not very large (say 5% or less), such an approach will not lose much power. If the fraction of incomplete cases is large, as is often the case when there are large numbers of variables in the model, then the case-deletion method will cause the loss of large amounts of information.

If the missing data are MAR, where the missingness depends on the observed variables, omitting incomplete cases from the analysis may lead to bias. Linear regression will yield a biased estimate of the effect of regressor X on outcome Y if the probability that X is missing depends on the value of Y. However, as long as the probability that X is missing depends only on the values of other regressors, not on Y, the estimated association between X and Y will still be unbiased. With logistic regression, estimates will be consistent as long as the probability of missing data on X depends only on the value of Y, or only on the values of the other covariates, but not both. Thus case deletion is reasonably robust but is not recommended when statistical power is a concern.

Pairwise deletion Linear regression estimates can be derived using the sample means and covariance matrix only. With pairwise deletion, each summary statistic (e.g., covariance) is estimated using all available cases for that statistic. Although this approach sounds as though it would be more efficient than case deletion, this is true only when the correlations among variables are low. Furthermore, pairwise deletion only yields consistent estimates under MCAR, not MAR, and in either case, the standard errors and test statistics will be biased. Thus pairwise deletion is not a recommended strategy.

Dummy variable adjustment (missing-indicator method) With this method, for each regressor X that has missing data, a corresponding dummy (0-1) variable is created and included in the regression along with X. The dummy variable is set to equal 1 if the value of the corresponding regressor is missing, and 0 otherwise. The missing values for the regressor are reset to any constant value k . The choice of the value for k (e.g., the mean of the regressor, or even zero) does not affect the estimated coefficient on the regressor (X), only the estimated coefficient on the dummy variable (D). A variant of this approach is used with categorical variables (e.g., race), where a separate dummy variable is constructed for missing values (e.g., the regression might include dummy variables for white, black, Asian, other race and "missing race"). Although

dummy variable adjustment has strong intuitive appeal and is very commonly used, it unfortunately yields biased estimates even when the data are MCAR. This technique should therefore be avoided.

Marginal mean imputation With marginal mean imputation, each missing value is replaced by the observed mean for that variable. This method preserves the observed sample means, but it dampens the relationship between variables and leads to an underestimate of the standard errors and p-values.

Conditional mean imputation A variant on marginal mean imputation is to use regression-adjusted means. Using the cases with complete data, the variable with missing data (X) is regressed on the other covariates. Using the resulting parameter estimates, predicted values for X are then generated and substituted for the missing values. In contrast to marginal mean imputation, use of conditional mean imputation generally will lead to artificially inflated correlations among variables. It is possible to obtain consistent coefficient estimates using conditional mean imputation if the data are MCAR and the imputations are based only on the other covariates (not the dependent variable), but even in this case, the standard errors will be underestimated.

Hot-decking and cold-decking Hot-decking (Brick and Kalton, 1996) involves the replacement of a missing value with a donor value taken from another matched respondent in the same imputation class. For example, a respondent with missing income data might be assigned the value of income reported by another respondent with similar age, sex, race and education. “Cold-decking” is similar to hot-decking, with the key difference that hot-decking uses donor values from other observations within the same dataset, whereas cold-decking uses donor values obtained externally, e.g., from other datasets with comparable variables, through the use of expert panels, or by assuming extreme values in order to test sensitivity of the findings. Hot-decking is very commonly used with large public use survey data, such as those issued by the Census Bureau.

With hot-decking, “recipient” observations can also be matched with “donor” observations through regression modeling rather than simple stratification, e.g., using propensity score matching. As an example of the latter, to impute missing values for income, one could estimate a logit model for the probability that the observation has a missing value for income, using other covariates (e.g., age, gender, race, education) as predictors in this model. Based on the parameter estimates from this initial model, the predicted probability of having a missing value for income is then constructed. Observations with missing data for income are then randomly matched to observations with non-missing data for income within groups that have similar predicted probabilities of missing income; a key decision is the choice of bandwidth around these predicted probabilities for purposes of defining groups for matching purposes. The value of income for the donor observation is then assigned to the recipient observation to which it was matched. An advantage of propensity score matching over simple stratification is that respondents can be “balanced” on many characteristics without running into sample size problems.

The hot-decking method improves upon mean imputation, although if only one imputation is performed, then it still has the problem of underestimating standard errors by not fully reflecting the random variation in the data. Importantly, hot-decking can be used in the context of multiple imputation as well (e.g., using the approximate Bayesian bootstrap or predictive mean matching), although the procedure involves extra steps to ensure that sufficient “noise” is added to the imputation process to get an estimate of the variability.

STATA has a user-written module to perform hot-deck imputation with propensity score matching. This module can be downloaded as a STATA-released ado file (STB51/sg116) by typing the following lines into the STATA editor window:

net from <http://www.mrc-bsu.cam.ac.uk/personal/adrian>
net install hotdeck

A full description of this module is recorded in a PDF file, which can be found at <http://www.ats.ucla.edu/stat/stata/stbpdf/hotdeck.pdf>.

Maximum likelihood (ML)

Maximum likelihood estimates are those estimates that, if true, maximize the likelihood that the observed sample is the one actually drawn. A simple example is that of estimating the probability of getting heads when flipping a weighted coin: If 10 flips of the coin yield 7 tails and 3 heads, then the maximum likelihood estimate of the probability of heads is .3. Choosing .3 as the estimated probability of heads maximizes the chance that we would actually observe 3 out of 10 coin flips to be heads. Maximum likelihood estimators have the advantage of being both “consistent” (in large samples, the estimates will be approximately unbiased) and “efficient” (no other consistent estimator yields smaller true standard errors).

In contrast to conventional methods, maximum likelihood treats missing data as an additional source of variability to be averaged over, rather than something to be eliminated (e.g., by deleting the observation or imputing the value). This is achieved by summing the “likelihood function” over all possible values for the missing data. For most applications, derivation of the maximum likelihood estimates requires an iterative process, since the estimates cannot be calculated through an explicit formula.

Maximum likelihood methods work best when the missing data are monotonic. Monotone missing patterns means that the data can be ordered so that everybody who is missing X_1 must also be missing X_2, \dots, X_n (where X_n is the last observation), all of those missing X_2 must also be missing X_3, \dots, X_n , and so forth. Monotone missing data patterns happen often when the survey is not completed due to lack of time, or when the respondent got tired or did not want to continue, so the respondent missed the last few questions of the survey. Although this is likely to happen in many surveys, there are also likely to be some non-monotonically missing values as well, i.e., if people completed the interview but chose not to respond to certain questions.

There are several maximum likelihood techniques that can be used to deal with missing data: linear models with normally distributed data, the expectation-maximization (EM) algorithm, and direct (or “raw”) maximum likelihood estimation. The first of these methods requires strong assumptions, i.e., that the variables can be modeled as a system of linear regressions with multivariate normal error terms. This distributional assumption is innocuous for variables without missing data, and even for variables with missing data, the results are often fairly robust to the assumption of normality. However, the ability to estimate models of this type is limited unless monotonicity holds.

The EM algorithm can deal with general (non-monotonic) missing data patterns and it also has the advantage of being easy to use and available in conventional software packages. EM is based on the fact that if we knew the true values for the missing data, we could estimate the model parameters, and if we knew the model parameters, we could obtain unbiased predictions for the missing values. Thus EM uses an iterative process in which the missing values are first imputed based on initial assumptions about the parameter values, then the imputed data are used to update the parameter estimates, then the updated parameter estimates are used to re-impute the missing data, and so forth until the estimates converge. The primary limitation of EM is that the standard errors and hence p-values will be underestimated. A technique known as direct (or “raw”) maximum likelihood estimation does not have this problem, but is more difficult to implement. Given the current state of theory and software, maximum likelihood estimation is

a useful option for dealing with missing data only when linear or log-linear models can be assumed.

Multiple imputation (MI)

Overview of multiple imputation An alternative method for dealing with missing data is multiple imputation. To account for the random error in the values of the missing variables, m possible values are imputed for each missing value under some presumed model, e.g., a normal distribution. Then the m imputed data sets are analyzed separately. The estimates from the m separate statistical analyses (typically the estimated regression coefficients, standard errors, and p values) are then combined into a single set of results by assuming an implicit or explicit model for the sampling distribution of the data.

The proper use of multiple imputation depends on making the correct assumption about the distribution of the data. Sometimes we do not know the distribution of the underlying data, in which case we have to rely on distributional approximations and the robustness of the complete-data inference. Once we make a distributional approximation, rules for combining the multiple imputation results (for example, obtaining a single set of estimated odds ratios from multiple logistic regressions) by basing inferences on large-sample normal theory.

The number of imputations necessary depends on how much of the data are missing and how much efficiency is desired. For example, to achieve 97% efficiency, only $m=3$ imputations are necessary if 10% of the data are missing, but this number goes up to $m=10$ if 30% of the data are missing and $m=20$ if 70% of the data are missing. A commonly used rule of thumb is to perform $m=5$ imputations.

Multiple imputation should give results similar to those obtained using maximum likelihood. As with maximum likelihood, multiple imputation yields estimates that are consistent and efficient. It has the added advantage of being substantially more flexible, as it can be used with all types of data and models. Multiple imputation involves a large amount of computation, but with the aid of statistical software and available computing power, this does not pose a major problem, even in a large data set. Conventional software (SAS, S-Plus) is available for many MI applications.

Creation of multiply imputed datasets Although a model is required to generate the imputations, MI methods are less sensitive to the assumptions of the model than ML methods, as the model is used only to generate the imputations and not to estimate the parameters of interest. The most popular method for creating multiply imputed datasets is the multivariate normal model, as it is quite robust even when the data are non-normal. Essentially, the procedure for creating imputations under this model involves regressing each variable with missing data on all of the other variables of interest, creating predicted values based on the estimated regression coefficients, and adding an error term which is a random draw from the residual normal distribution of the variable.

This approach, however, does not take into account the fact that the regression coefficients in the imputation model are estimates and not true values, so are themselves subject to variability. To address this issue, one would ideally use parameter estimates that are random draws from a Bayesian posterior distribution, based on the actual and observed data. This can be accomplished by using the technique of data augmentation, also known as the Markov Chain Monte Carlo (MCMC) algorithm, which is a general method for finding a Bayesian posterior distribution. Similarly to the EM algorithm described earlier, MCMC alternately imputes the missing data and updates the unknown parameter values. With MCMC, however, the latter is done in a random fashion. Each new set of imputations allows the Bayesian posterior distribution for the parameter values (rather than the parameter estimates themselves) to be

updated, and then new parameter values for the imputation model are randomly drawn from this distribution in order to update the imputations.

One implication is that with MCMC, it is not the parameter values themselves that converge, but the *distribution* of the parameters. This raises the question of how many iterations to run and how convergence is determined. Since it is the distribution that converges rather than the estimates themselves, it is less straightforward to determine convergence. One rule of thumb is to perform EM estimation first and use at least the same number of iterations that it takes for the EM estimates to converge. This method is also advisable because the EM estimates provide useful starting values for the parameter estimates. A test of the serial correlation between MCMC iterations should also be performed. If the estimates are serially correlated, then the distribution has not yet converged. With a higher proportion of missing data, more iterations will be required.

The PROC MI procedure in SAS can create multiply imputed datasets based on a multivariate normal model, using regression, propensity score or MCMC methods. S-Plus also uses the MCMC method. The first two methods (regression and propensity score) are for normal monotone missing patterns only, not for randomly missing cases. However, the method used to create the imputations may not make much of a difference if the proportion of missing values is not that great.

Analysis of multiply imputed datasets After the multiple imputations are created, m plausible versions of the complete data exist, each of which are analyzed by standard complete-data methods. The results of the m analyses are then combined to produce a single inferential statement that includes uncertainty due to missing data. PROC MIANALYZE in SAS can then be used to combine standard regression estimates (linear and dichotomous logit) from the multiply imputed datasets. Other software packages can do the same thing, for example, NORM (Schafer, 1997) or a user-written macro within STATA.

The following summarizes multiple imputation inference for the univariate case (Rubin, 1987; Schafer, 1997). Suppose the data is under a multivariate normal model, and we draw m values for each missing item. With the m complete data sets, we do data analysis separately and get m sets of estimates of the parameter of interest, Q , and the associated variance, U (the square of the standard error). Then we can use the following formulae to get one final set of parameter estimate, its corresponding standard error, and the p value from the t test.

The average of the estimate of Q can be expressed as:

$$\bar{Q} = \frac{1}{m} \sum_{l=1}^m Q_l$$

The within-imputation variance is:

$$\bar{U} = \frac{1}{m} \sum_{l=1}^m U_l$$

The between-imputation variance is:

$$B = \frac{1}{m-1} \sum_{l=1}^m (Q_l - \bar{Q})^2$$

Then the total variance is:

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

The number of degrees of freedom is:

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

and the t test statistic is:

$$\frac{Q - \bar{Q}}{\sqrt{T}} \sim t_v$$

By using the approximation formula corresponding to the appropriate degrees of freedom, a t test statistic and a p value associated with testing whether the true parameter is significantly different from the parameter estimate can be calculated.

Correspondence between the imputation and estimation model For consistency of the final estimates, it is critical to include all of the variables in the imputation model that will eventually be included in the subsequent estimation model that is the focus of the analysis. In general, the imputation model must preserve the relationships between variables that will be investigated in the subsequent estimation model. For example, if a variable Y is imputed using a variable X_1 but not X_2 , and then Y is subsequently regressed on X_2 as well as X_1 , the estimated coefficient for X_2 would be biased towards zero. For the same reason, one must decide ahead of time whether interactions will be included in the final estimation model, since there are special procedures that must be followed to ensure that imputation does not dampen evidence of interaction effects in the estimation model.

The converse, however, is not true; one can use additional variables in the imputation model without necessarily including them in the final estimation model. Indeed, one strategy is to add as many variables as possible to the imputation model, to ensure that all variables of possible interest for the subsequent estimation model are included at the imputation stage. However, inclusion of additional predictors may actually worsen the imputations unless the extra predictors are sufficiently correlated with the variables to be imputed. PROC MI automatically uses all of the variables for which missing values are being imputed and also allows the inclusion of additional covariates that may be useful in predicting the missing data values.

Role of the dependent variable in multiple imputation Two issues arise with regard to the dependent variable in an analysis. The first is whether the dependent variable should be used to help impute independent variables. If deterministic imputation is used (e.g., plugging in a regression-adjusted mean), inclusion of the dependent variable in the imputation model will inflate the association of the dependent and independent variables in the estimation model. However, with multiple imputation, random error is introduced, thereby ensuring that the estimates are consistent. Leaving the dependent variable out of the imputation model would actually underestimate the coefficients on the imputed regressors in the estimation model. Therefore the dependent variable should be included in the imputation model for the covariates.

The second issue is whether the dependent variable itself should be imputed. The answer to this question is somewhat counterintuitive: If missing data occur only with the dependent variable, then the cases should simply be deleted, as is usually done in the literature. However, if there are missing data for both the dependent and independent variables, then the cases with missing data for the dependent variable should be retained, and the values imputed.

Multiple imputation with non-normally distributed variables Health services research often relies on categorical data, whereas the multiple imputation model assumes that all the variables in the model are multivariate normally distributed. Nonetheless, Belin et al. (2000) found that

the normal model produced more accurate predictions than the general location model, despite the fact that the latter is tailored to accommodate both categorical and continuous variables. This evidence suggests that it is still feasible to use the multivariate normal distribution model to multiply impute categorical variables. Other techniques, such as transforming skewed data (e.g., using a log) to perform the imputations and then retransforming the imputed values, can improve the performance of the normal model when imputing continuous but non-normal data.

Use of the multivariate normal model results in continuous imputed values for the binary variable. Thus it is necessary to set up the lower and upper limits of the values, and round the imputed value to the nearest integer. For example, suppose there are missing values for an indicator for whether the person reports poor health. Despite the dichotomous nature of this variable, a linear regression would be used for estimation in the MI procedure, resulting in a (continuous) predicted probability of poor health for each individual. If this predicted probability is above a certain cutoff (e.g., .5), then the individual's imputed value is set equal to one, and if it is below, then the imputed value is set equal to zero. For categorical variables with more than two categories, such as race, the dichotomous indicators for each category are predicted separately and then the individual is assigned to whichever of the categories has the highest predicted probability.

In the SAS PROC MI procedure, the values are automatically rounded when the lower and upper limits of the imputed values are set. Other macros/software programs are written specifically for categorical variables using the log linear model, for example, the CAT library in S-PLUS, etc. S-Plus handles categorical variables better than SAS and can deal with normal, categorical, ordered categorical and mixed normal and categorical missing value cases. Note, however, that imputation of categorical variables involves the estimation of more parameters than with continuous variables, so it requires a larger sample size.

Degrees of freedom Another important limitation of multiple imputation is that it may require more degrees of freedom than we have. For example, one cannot multiply impute more variables than the number of observations in the sample, since multiple imputation requires the estimation of a system of simultaneous equations, one per variable to be imputed. This limitation does not hold for every missing data method; for example, marginal mean imputation or hot-decking allow the researcher to impute all variables in a dataset.

Non-standard cases Another limitation is that there is no good method to handle the case in which the complete-data inference is not based on large-sample normal theory. For example, the parameters of interest from logistic regressions are usually relative risks². A single combined estimate of the relative risk for a particular covariate can be obtained by calculating the relative risk from each logistic regression and taking an average of the m estimates. However, calculation of the combined confidence interval around this estimate is more complicated. For each of the m iterations, the confidence intervals must be bootstrapped, since no closed-form solution exists. Unless the bootstrap estimates can be reasonably assumed to follow a normal distribution, the confidence intervals must be derived empirically. With 1000 bootstrap repetitions and a 5% type I error in a two-tailed test, one would rank the estimates and then choose the 26th estimate as the lower confidence limit and the 975th estimate as the upper confidence limit. In this situation, there are no variance estimates to use in the formulae listed above. One option is to use the smallest of the lower confidence bounds and the largest of the upper confidence bounds from the m confidence intervals calculated from the datasets. While this approach produces a confidence interval with at least $(1-\alpha)\%$ probability of coverage of the true value of the relative risk, it is likely to be conservative.

² Odds ratios are often interpreted as though they were relative risks, since they are sometimes a good approximation, but it is usually preferable to just directly calculate the relative risks themselves.