Chapter 20

Missing Data

Paul D. Allison

20.1. Introduction

Perhaps the most universal dilemma in statistics is what to do about missing data. Virtually every data set of at least moderate size has some missing data, usually enough to cause serious concern about what methods should be used. The good news is that the last twenty-five years have witnessed a revolution in methods for handling missing data, to the point where there is little overlap between this chapter and the missing data chapter in the 1983 edition of this *handbook*. The methods that have been developed in the interim have much better statistical properties than traditional methods, while at the same time relying on weaker assumptions.

The bad news is that these superior methods have not been widely adopted by practicing researchers. The most likely reason is ignorance. Many researchers have barely even heard of modern methods for handling missing data. And if they have heard of them, they have little idea how to go about implementing them. The other likely reason is difficulty. Modern methods can take considerably more time and effort, especially with regard to start-up costs. Nevertheless, with the development of better software, these methods are getting easier to use every year.

Three broad classes of missing data methods have good statistical properties: maximum likelihood (ML), multiple imputation (MI), and inverse probability weighting. ML and MI can handle a wide array of applications, and many commercial software packages implement some version of these methods. As of this writing, inverse probability weighting is much more limited in its applications, and easy-to-use software is not readily available. For that reason, I will not discuss inverse probability weighting in this chapter. However, it should be noted that this method may be more robust than ML and MI to certain kinds of misspecification (Robins, Rotnitzky, & Zhao, 1995; Robins & Rotnitzky, 1995; Scharfstein, Rotnitzky, & Robins, 1999).

Handbook of Survey Research, Second Edition
Copyright © 2010 by Emerald Group Publishing Limited
All rights of reproduction in any form reserved
ISBN: 978-1-84855-224-1

This chapter briefly reviews the strengths (few) and weaknesses (many) of conventional methods for handling missing data. I then examine ML and MI in some detail, emphasizing their conceptual foundations and illustrating their implementation using currently available software. Space does not permit a very rigorous or technical treatment of these methods; see Allison (2001) for a more extended introduction to them.

20.2. Example

To illustrate the various methods for handling missing data, I use data from the National Survey of Families and Households (NSFH) (Sweet & Bumpass, 2002). The NSFH is a national probability sample survey of 13,007 adults age 19 and older who were interviewed initially in 1987–1988. They were reinterviewed in 1992–1994 and again in 2001–2002. The analysis sample used here ($N = 3622$) consists of couples who met the following criteria: (a) they were married at wave 1, (b) both spouses completed the initial interview, and (c) at least one spouse was reinterviewed at a later wave.

Our analytical goal is to estimate a logistic regression model in which the dependent variable is whether or not the couple divorced between the initial interview and a later wave. There were 747 couples who divorced (20 percent), and no missing data on the dependent variable. Table 20.1 lists the variables used as predictors in the model, along with the number of nonmissing cases and the minimum and maximum values for each variable. All the variables with a minimum value of 0 and a maximum of 1 are dummy variables. The variables FBOD and MBOD are multiple-item scales that measure, respectively, the wife's and husband's assessment of whether they would be ‘‘better off divorced.’’

Only one predictor variable (CLT6) has no missing data. For the other variables, the proportion of cases with missing data is generally small, with a maximum of 10 percent missing for FRELG. Nonetheless, 1182 cases (33 percent) have missing data on at least one of the 14 variables, and all of them would by definition be lost to the analysis under listwise deletion (complete case analysis).

The first panel of Table 20.2 presents estimates for a conventional logistic regression using listwise deletion. Couples are more likely to divorce if either the husband or the wife was previously divorced, if either spouse assessed the likelihood of separation to be high, or if either spouse felt that they would be ‘‘better off divorced.’’ Divorce is more likely if there are children under six or if the husband is not Catholic.

20.3. Assumptions

When we think of missing data, we usually envision situations in which a variable has some ‘‘real’’ value, but we simply do not observe it for one reason or another. For

Table 20.1: Predictor variables for logistic regression.

Variable	**Label**	**Non-missing cases**	**Minimum**	**Maximum**
MEDUC	Husband's years of education	3593	0	20.0000000
FEDUC	Wife's years of education	3601	0	20.0000000
FGENID	Wife gender ideology scale. High is egalitarian	3444	− 3.5322108	2.6024005
MPRVDV	Husband previously divorced	3603	0	1.0000000
FPRVDV	Wife previously divorced	3594	0	1.0000000
CLT6	Presence of children under 6	3622	0	1.0000000
FPRBLOW	Wife assessment of prob separate 1 = low	3379	0	1.0000000
MPRBLOW	Husband assessment of prob separate 1 = low	3386	0	1.0000000
FBOD	Wife better off divorced scale	3326	− 1.9715052	4.1361923
MBOD	Husband better off divorced scale	3346	− 2.2109966	3.7527237
FCATH	Wife catholic	3578	0	1.0000000
MCATH	Husband catholic	3572	0	1.0000000
FRELG	Wife how often attends religious services	3251	1.0000000	7.0000000
MRELG	Husband how often attends religious services	3577	1.0000000	7.0000000

example, everyone has a real age but some people choose not to report it. In many circumstances, however, a variable is not defined or has no meaning for some people. For example, "marital happiness" has no meaning for people who are not married. Likewise, "job satisfaction" is undefined for those who are not employed. Such variables are usually accommodated by skip patterns in survey instruments.

The methods discussed in this chapter are primarily designed for the first situation in which a real value happens to be missing, although some authorities (Schafer, 1997) endorse multiple imputation even for variables that are not defined for some individuals. For the latter situation in which there is no real value to be imputed, different methods may be appropriate. One such method, known as dummy variable adjustment, is discussed in the next section.

There is no end to the list of possible reasons why real data might be missing. People may refuse to answer questions or give incoherent answers. In a longitudinal

Table 20.2: Logistic regression predicting divorce, using various methods for missing data.

Parameter	Listwise deletion		Multiple imputation (MCMC)[a]		Multiple imputation (SGR)[b]		Maximum likelihood (MVN)[c]		Maximum likelihood (MVN & Logit)[d]	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
MEDUC	–.022	.026	–.037	.020	–.037	.020	–.038	.020	–.037	.020
FEDUC	–.019	.029	–.017	.023	–.015	.023	–.014	.023	–.015	.023
FGENID	.093	.060	.140**	.051	.137**	.052	.135**	.050	.135**	.050
MPRVDV	.497***	.138	.518***	.111	.517***	.111	.519***	.110	.520***	.111
FPRVDV	.367*	.143	.396***	.114	.394***	.115	.393***	.114	.393***	.114
CLT6	.252*	.111	.280**	.089	.277**	.090	.276**	.089	.276**	.090
FPRBLOW	–.953***	.168	–1.034***	.138	–1.041***	.136	–1.011***	.038	–1.014***	.138
MPRBLOW	–.551**	.176	–.458***	.137	–.423***	.134	–.444***	.140	–.442***	.140
FBOD	.250***	.061	.222***	.050	.232***	.051	.232***	.050	.228***	.050
MBOD	.156**	.060	.170***	.049	.161***	.048	.168***	.049	.167***	.049
FCATH	.269	.155	.074	.128	.069	.130	.069	.129	.072	.129
MCATH	–.328*	.165	–.304*	.134	–.300*	.136	–.298*	.134	–.301*	.134
FRELG	–.018	.034	–.021	.027	–.020	.029	–.017	.028	–.015	.028
MRELG	–.066	.035	–.056*	.027	–.057*	.028	–.059*	.028	–.061*	.028

Note: *$.01<p<.05$; **$.001<p<.01$; ***$p<.001$.

[a] Markov Chain Monte Carlo imputations were obtained under a multivariate normal model.

[b] Sequential generalized regression imputations used logistic regression for imputation of all dichotomous variables.

[c] Assumed that all predictors had a multivariate normal distribution.

[d] Assumed that dichotomous predictors with missing data were logistic regression functions of other variables; all other variables were assumed to be multivariate normal.

survey, people may respond to earlier waves but not to later ones. Interviewers may forget to ask some questions. Administrative data may be lost. With respect to missing data methods, however, exactly why the data are missing does not really matter. What does matter is whether the data are missing completely at random, missing at random, or not missing at random.

20.3.1. Missing Completely at Random

Suppose that only one variable Y has missing data, and that another set of variables, represented by the vector X, is always observed. The data are *missing completely at random* (MCAR) if the probability that Y is missing does not depend on X or on Y itself (Rubin, 1976). To represent this mathematically, let R be a "response" indicator having a value of 1 if Y is missing and 0 if Y is observed. We then have

$$\Pr(R = 1|X, Y) = \Pr(R = 1)$$

A natural question to ask at this point is, what variables can be or should be in the X vector? The answer is quite simple. The only variables that *must* be in X are those that are part of the model to be estimated. Suppose, for example, that we seek only to estimate the mean income for some population, and 20% of the cases are missing data on income. In that case, we need not consider any X variables for the MCAR condition. The only relevant question is whether the probability that income is missing depends on income itself, for example, whether people with high income are less likely to report their income. On the other hand, if our goal is to estimate the correlation between income and years of schooling, the MCAR condition requires that missingness on income not depend on either income or years of schooling.

How can we test the MCAR assumption? Testing for whether missingness on Y depends on some observed variable X is easy. For example, we can test whether missingness on income depends on gender by testing whether the proportions of men and women who report their income differ. More generally, we could run a logistic regression in which the dependent variable is the response indicator R and the independent variables are all X variables in the model to be estimated. Significant coefficients would suggest a violation of MCAR.

On the other hand, it is not so easy to test the other part of MCAR, that missingness on Y does not depend on Y itself. For example, the only way to test whether people with high incomes are less likely to report their incomes is to find some other measure of income (e.g., tax records) that has no missing data. But this is rarely possible.

The MCAR assumption is very strong, and is unlikely to be completely satisfied unless data are *missing by design* (Graham, Hofer, & MacKinnon, 1996). A well-known example is the General Social Survey (Davis, Smith, & Marsden, 1972–2006), which typically has three different "ballots" containing different sets of questions on social and political attitudes and behaviors. Each ballot is administered to a random two-thirds of the respondents, and all pairs of items are administered to some respondents. The goal is to cover a wider range of topics without increasing interview time. Any data that are missing by this design are missing completely at random.

This example illustrates another key point about MCAR. The assumption is not violated if the probability that one variable is missing is related to whether another variable is missing. In the General Social Survey design, the questions in each ballot would be either all observed or all missing for any respondent.

20.3.2. Missing at Random

A considerably weaker (but still strong) assumption is that data are *missing at random* (MAR). Again, this is most easily defined in the case where only a single variable Y has missing data, and another set of variables X has no missing data. We say that data on Y are missing at random if the probability that Y is missing does not depend on Y, once we control for X. In symbols, we have

$$\Pr(R = 1|X, Y) = \Pr(R = 1|X)$$

where R is the response indicator. Thus, MAR allows for missingness on Y to depend on other variables that are observed. It just cannot depend on Y itself (after adjusting for the observed variables).

As with MCAR, the only variables that *must* go into X are the variables in the model to be estimated. But under MAR, there can be substantial gains from including other variables as well. Suppose, for example, that we believe that people with high income are less likely to report their income. That would violate both MCAR and MAR. However, by adjusting for other variables that are correlated with income — for example, education, occupation, gender, age, mean income in zipcode — we may be able to greatly reduce the dependence of missingness of income on income itself.

The MAR condition is often equated with the phrase "ignorability of the missing data mechanism," but ignorability is actually somewhat stronger (Rubin, 1976). The "missing data mechanism" is simply the equation that expresses the probability of missingness as a function of Y and X. Ignorability requires that the data be MAR, *and* that the parameters that govern the missing data mechanism be functionally distinct from the parameters of the model that govern the data itself. This is a rather technical condition that is unlikely to be violated and, hence, will not be discussed in any detail. As the word suggests, ignorability implies that we can obtain optimal estimates of the data parameters without having to model the missing data mechanism. Henceforth, I will use the terms MAR and ignorable interchangeably.

20.3.3. Not Missing at Random

We say that data are *not missing at random* (NMAR) if the MAR assumption is violated, that is, if the probability that Y is missing depends on Y itself, after adjusting for X. There are often strong reasons for suspecting that the data are

NMAR, for example, people who have been arrested may be less likely to report their arrest status. However, as previously noted, the data contain no information for testing such suspicions. If the data are truly NMAR (and, thus, missingness is not ignorable), then the missing data mechanism must be modeled as part of the estimation process in order to produce unbiased parameter estimates. This is not straightforward because one can always specify an infinite number of different models for the missing data mechanism. Nothing in the data will indicate which of these models is correct. And, unfortunately, results may be highly sensitive to the choice of model.

Because of these difficulties, most software packages that implement ML or MI assume ignorability (and, hence, MAR). Both ML and MI can produce optimal estimates in the NMAR case (as discussed in Section 20.6) with a correct model for the missing data mechanism, but it is difficult to have confidence that any given model is correct.

20.4. Conventional Methods

Before proceeding to ML and MI, let us briefly review more conventional methods for handling missing data, along with their strengths and weaknesses.

20.4.1. Listwise Deletion

The most common method for handling missing data is listwise deletion, also known as complete case analysis. This method simply deletes observations that have missing data on any variables in the model of interest. Only complete cases are used.

Listwise deletion has two big and obvious attractions: it is easy and can be used with any statistical method. Furthermore, if the data are MCAR, listwise deletion will not introduce any bias into estimates. That is because, under MCAR, the subsample of complete cases is effectively a simple random sample from the original sample, and it is well known that simple random sampling does not introduce bias (see, e.g., Frankel, this volume). Last, and quite important, listwise deletion produces estimated standard errors that consistently estimate the true standard errors. Thus, unlike conventional imputation methods, listwise deletion is "honest": it does not assume that one has more or better data than are actually available.

The obvious downside of listwise deletion is that, quite often, it discards a great deal of potentially useful information. As a consequence, the true standard errors may be much higher than necessary, implying unnecessarily wide confidence intervals and high *p*-values. A second undesirable feature of listwise deletion is that parameter estimates *may* be biased if the data are MAR but not MCAR. For example, if men are less likely to report income than women, estimates of mean income for the whole population are likely to be biased downward.

Violation of MCAR does not *always* result in biased estimates under listwise deletion, however. In fact, when predictor variables in regression analysis (either linear or logistic) have missing data, listwise deletion yields unbiased estimates of coefficients even when the data are not missing at random (Little, 1992). Thus, even if high income people are less likely to report their income, coefficients for income as a predictor are not biased by listwise deletion. For a proof, details and a caveat, see Allison (2001). Of course, deletion of what may be a large number of cases may still result in a loss of power.

20.4.2. Pairwise Deletion

Pairwise deletion, also known as available case analysis, is a popular method for handling missing data when estimating linear regression and other linear models. It rests on the fact that, for a wide class of linear models, the "minimal sufficient statistics" are the means, variances, and covariances. That implies that the parameter estimates can be computed using only these statistics, without any loss of information. In pairwise deletion, the means and variances are estimated using all nonmissing cases for each variable. Each covariance is estimated using all cases with data present on both variables. Once the means, variances, and covariances have been estimated, they are simply plugged into standard formulas for the linear model parameters.

This method utilizes all the data and deletes no cases. It would seem that pairwise deletion ought to perform much better than listwise deletion, and it is easily shown that under MCAR, pairwise deletion provides consistent (and, hence, approximately unbiased) parameter estimates (Glasser, 1964). But, like listwise deletion, pairwise deletion can introduce bias if the data are MAR but not MCAR. Also, pairwise deletion sometimes breaks down entirely because the correlation matrix may have patterns that are simply not possible with complete data.

Last, and perhaps most important, with pairwise deletion it is difficult to obtain accurate standard error estimates, *p*-values, and confidence intervals. When a pairwise-deleted covariance matrix is used within linear modeling software, one must specify a sample size to get standard errors and *p*-values. But what sample size is appropriate? Certainly not the original sample size: that would not take into account the fact that data, perhaps many data, are missing. But the *listwise*-deletion sample size would be too small because pairwise deletion uses more of the data. The fact is that no single sample size will give the correct standard errors for all the parameter estimates under pairwise deletion.

20.4.3. Dummy Variable Adjustment

Another popular approach to handling missing data on predictors in regression analysis is dummy variable adjustment (Cohen & Cohen, 1985). Its mechanics are simple and intuitive. Suppose one predictor is income, with missing data on, say, 30% of the cases. The analyst creates a dummy variable with a value of 1 for people who are missing income and 0 for people who are not. Next, for people with missing income, one ‘‘imputes’’ some constant value — for example, the mean for nonmissing cases. Then a regression is estimated with both income and the dummy variable as predictors, along with any other predictors. Such dummy variable adjustments can be made for many (even all) predictor variables.

The appeal of this method is that it deletes no cases, and incorporates all available information into the regression model. But Jones (1996) proved that dummy variable adjustment yields biased parameter estimates even when the data are MCAR, which pretty much rules it out. Jones also demonstrated that a related method for nominal predictors produces biased estimates. That method treats missing cases for a categorical variable simply as another category, creating an additional dummy variable for that category.

Despite this apparently fatal flaw, dummy variable adjustment can be useful in two situations. Suppose, first, that the primary goal is to generate good predictions, and further that missing data are anticipated for at least some of the out-of-sample cases for which predictions are desired. Unbiased parameter estimation is not essential for good predictive modeling, and the coefficients for the dummy variables will yield predictions even for cases with missing data.

Second, suppose that data are ‘‘missing’’ because a variable is undefined for some subset of the sample, for example, marital satisfaction for unmarried persons. Jones’ proof presumes that a missing datum has some real value that is simply not observed. It is easy to show, however, that the dummy variable adjustment method leads to unbiased estimates under a simple but plausible model for the situation in which the missing item is undefined (proof available on request).

20.4.4. Imputation

The basic principle of imputation is to generate plausible values for the missing values, and then do analyses as if no data were missing. There are many ways to do this. One of the simplest is to replace the missing values by sample means calculated for the nonmissing cases. It is well known that mean substitution produces biased estimates for most parameters, even under MCAR (Haitovsky, 1968). Better results can be obtained by using linear regression to generate imputed values (sometimes known as conditional mean imputation).

All conventional imputation methods suffer from two serious problems. First, variances tend to be underestimated, leading to biases in other parameters (like correlations and regression coefficients) that depend on variances. Mean substitution, for example, replaces the presumably different missing values with a single value, thereby reducing the variance. Regression-based imputation also understates variances, although to a lesser degree.

The second problem is equally serious. Standard data analysis software cannot distinguish imputed data from real data. In particular, standard error calculations presume that all data are real. The inherent uncertainty and sampling variability in the imputed values is not taken into account. As a result, reported standard errors are too low, sometimes much too low — leading, of course, to confidence intervals that are too narrow and *p*-values that are too low. In short, conventional imputation methods are fundamentally dishonest. In the next section, we will see how multiple imputation solves both of these problems.

20.5. Multiple Imputation

The three basic steps to multiple imputation are as follows:

1. Introduce random variation into the imputation process, and generate several data sets, each with slightly different imputed values.
2. Perform an analysis on each of the data sets.
3. Combine the results into a single set of parameter estimates, standard errors, and test statistics.

The first step is by far the most complicated, and there are many different ways to do it. One popular method uses linear regression imputation. Suppose a data set has three variables, X, Y, and Z. Suppose X and Y are fully observed, but Z has missing data for, say, 20% of the cases. To impute the missing values for Z, a regression of Z on X and Y for the cases with no missing data yields the imputation equation

$$\hat{Z} = b_0 + b_1 X + b_2 Y$$

Conventional imputation would simply plug in values of X and Y for the cases with missing data and calculate predicted values of Z. But, as noted previously, such imputed values have too small a variance. To correct this problem, we instead use the imputation equation

$$\hat{Z} = b_0 + b_1 X + b_2 Y + sE,$$

where E is a random draw from a standard normal distribution (with a mean of 0 and a standard deviation of 1) and s is the estimated standard deviation of the error term in the regression (the root mean squared error). Adding this random draw raises the variance of the imputed values to approximately what it should be and, hence, avoids the biases that usually occur with conventional imputation.

If parameter bias were the only issue, imputation of a single data set with random draws would be sufficient. Standard error estimates would still be too low, however, because conventional software cannot take account of the fact that some data are imputed. Moreover, the resulting parameter estimates would not be fully efficient (in the statistical sense), because the added random variation introduces additional sampling variability.

The solution is to produce several data sets, each with different imputed values based on different random draws of E. The desired model is estimated on each data set, and the parameter estimates are simply averaged across the multiple runs. This yields much more stable parameter estimates that approach full efficiency.

With multiple data sets we can also solve the standard error problem, by calculating the variance of each parameter estimate across the several data sets. This "between" variance is an estimate of the additional sampling variability produced by the imputation process. The "within" variance is the mean of the squared standard errors from the separate analyses of the several data sets. The standard error adjusted for imputation is the square root of the sum of the within and between variances (applying a small correction factor to the latter). The formula (Rubin, 1987) is as follows:

$$\sqrt{\frac{1}{M}\sum_{k=1}^{M} s_k^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{k=1}^{M}(a_k-\bar{a})^2}$$

In this formula, M is the number of data sets, s_k is the standard error in the kth data set, a_k is the parameter estimate in the kth data set, and $\bar{a}$ is the mean of the parameter estimates. The factor $(1 + (1/M))$ corrects for the fact that the number of data sets is finite.

How many data sets are needed? With moderate amounts of missing data, five are usually enough to produce parameter estimates that are more than 90 percent efficient. More may be needed for good estimates of standard errors and associated statistics, however, especially when the fraction of missing data is large. I discuss this issue in more detail later.

20.5.1. Complications

This method for multiple imputation is pretty good, but it still produces standard errors that are little too low, because it does not account for the fact that the parameters in the imputation equation (b_0, b_1, b_2, and s) are only estimates with their own sampling variability. This can be rectified by drawing the imputation parameter values used to create each data set at random from an appropriate distribution. To use Bayesian terminology, these values must be random draws from the posterior distribution of the imputation parameters.

Of course, Bayesian inference requires a prior distribution reflecting prior beliefs about the parameters. In practice, however, multiple imputation almost always uses noninformative priors that have little or no content. One common choice is the Jeffreys prior, which implies that the posterior distribution for s is based on a chi-square distribution. The posterior distribution for the regression coefficients (conditional on s) is multivariate normal, with means given by the estimated values b_0, b_1, b_2, and a covariance matrix given by the estimated covariance matrix of those coefficients. For details, see Schafer (1997). The imputations for each data set are based on a separate random draw from this posterior distribution. Using a different set of imputation parameters for each data set induces additional variability into the imputed values across data sets, leading to larger standard errors using the formula above.

Now we have a *very* good multiple imputation method, at least when only one variable has missing data. Things become more difficult when two or more variables have missing data (unless the missing data pattern is monotonic, which is unusual). The problem arises when data are missing on one or more of the potential predictors, X and Y, used in imputing Z. Then no regression that we can actually estimate utilizes all of the available information about the relationships among the variables. Iterative methods of imputation are necessary to solve this problem.

The Markov chain Monte Carlo (MCMC) method, widely used for Bayesian inference (Schafer, 1997), is the most popular iterative algorithm for multiple imputation. For linear regression imputation, one MCMC iteration proceeds roughly as follows. We begin with some reasonable starting values for the means, variances, and covariances among a given set of variables. For example, these could be obtained by listwise or pairwise deletion. We divide the sample into subsamples, each having the same missing data pattern (i.e., the same set of variables present and missing). For each missing data pattern, we use the starting values to construct linear regressions for imputing the missing data, using all the observed variables in that pattern as predictors. We then impute the missing values, making random draws from the simulated error distribution as described above, which results in a single "completed" data set. Using this data set with missing data imputed, we recalculate the means, variances, and covariances, and then make a random draw from the posterior distribution of these parameters. Finally, we use these drawn parameter values to update the linear regression equations needed for imputation.

This process is usually repeated many times. For example, the SAS implementation (www.sas.com) runs 200 iterations of the algorithm before selecting the first completed data set, and then allows 100 iterations between each successive data set. So producing the default number of five data sets requires 600 iterations (each of which generates a data set). Why so many iterations? The first 200 ("burn-in") iterations are designed to ensure that the algorithm has converged to the correct posterior distribution. Then allowing 100 iterations between successive data sets gives us confidence that the imputed values in the different data sets are statistically independent. In my opinion, these numbers are far more than enough for the vast majority of applications.

Many software packages implement the method just described. The first was a stand-alone package called NORM (www.stat.psu.edu/~jls), which has also been incorporated into Splus (www.insightful.com). The SAS procedure MI is essentially a NORM clone. And Stata (www.stata.com) now has a multiple imputation command (mi) that uses this method.

If all assumptions are satisfied, the MCMC method produces parameter estimates that are consistent (and hence approximately unbiased in large samples), asymptotically normal, and almost fully efficient. Full efficiency would require an infinite number of data sets, but a relatively small number gets very close. The key assumptions are, first, that the data are missing at random (although multiple imputation methods exist for the NMAR case). Second, linear regression imputation implicitly assumes that the variables have a multivariate normal distribution.

Although this is a strong assumption, the MCMC method seems to work well even when it is clearly violated. But more on that later.

20.5.2. Example

We illustrate the MCMC linear imputation method for the NSFH data using the MI procedure in SAS. The following code imputes the data:

```
proc mi data = couple out = divout;
  var div m1educ f1educ f1genid m1prvdv f1prvdv
    clt6 fprblow mprblow fbod mbod
    fcatholic mcatholic frelg mrelg;
run;
```

The first line specifies the input SAS data set ("couple"), and the output SAS data set ("divout"), which contains all of the imputed data sets. The remaining lines specify a list of variables. At a minimum, the list should include all variables in the model to be estimated to ensure that the imputed values accurately reflect all the relationships among the variables. In particular, it is essential to include the dependent variable ("div" in this case) as a predictor for the other variables. I do not generally recommend imputing the dependent variable itself, however, so I usually delete cases missing data on the dependent variable before doing the imputation (with one important exception noted in the next paragraph).

In addition, it is often desirable to include *auxiliary* variables that are not intended to be in the final model. Ideally, these variables would be at least moderately correlated with variables in the model that have missing data. By improving the reliability of the imputations, using auxiliary variables can substantially reduce standard errors in the model to be estimated. If auxiliary variables are also associated with whether or not variables are missing, including them can reduce bias as well. The one situation in which it is useful to impute the dependent variable occurs when it is strongly associated with an auxiliary variable. This situation can be especially relevant to longitudinal studies, where missing data for a variable are often well predicted using the same variable measured at a different point in time as an auxiliary variable.

By default, the MI procedure produces five completed data sets. These are "stacked" into a single SAS data set ("divout"), along with a new variable "_imputation_" with values 1 through 5 to distinguish the different data sets.

Next, we estimate the logistic regression model on each of the five data sets using the "by" statement in SAS:

```
proc logistic data = divout outest = a covout;
  model div(desc) = m1educ feduc fgenid mprvdv fprvdv clt6 fprblow
    mprblow fbod mbod fcatholic mcatholic frelg mrelg;
  by _imputation_;
run;
```

This produces five sets of regression output. The "outest = a" option writes the regression coefficients to a SAS data set named "a." The "covout" option includes the covariance matrix of the coefficient estimates in that data set. Then the companion procedure MIANALYZE combines the results:

```
proc mianalyze data = a;
  modeleffects intercept m1educ feduc fgenid mprvdv fprvdv
    clt6 fprblow mprblow fbod mbod fcatholic mcatholic frelg mrelg;
run;
```

Using data set "a," MIANALYZE calculates the means of the parameter estimates across the five regression runs, and the standard errors using the formula above. The second panel of Table 20.2 above shows the results.

The coefficients estimated by multiple imputation are fairly close to those obtained with listwise deletion, except for the coefficient of FGENID, which is 50% larger with MI. The most striking thing about the multiple imputation results, however, is that all standard errors are lower than the corresponding ones for listwise deletion — exactly what we hope for. As a consequence, two coefficients (for FGENID and MRELG) that were not significant at the .05 level using listwise deletion are now significant using MI. And *p*-values for several other variables have declined as well. Typically, standard errors decrease the most for variables that have the least missing data because we are adding real data for these variables, data that listwise deletion does not use.

Because random draws are made at two critical points of the MCMC algorithm, the MI results in Table 20.2 would all change if we ran the procedure again. Sometimes the variability in results from one run to another can be large enough to change conclusions. As explained below, however, increasing the number of data sets can reduce this variability as much as desired.

20.5.3. Non-Normality

The MI method just illustrated assumes that the data have a multivariate normal distribution, implying that optimal imputation can be based on linear models. In the example, however, most of the imputed variables are dichotomous and of those that are not, some (e.g., FRELG and MRELG) are highly skewed. It is natural to question whether linear imputation models are appropriate for such variables. When variables have no missing data, the normality assumption is of little or no consequence. And a good deal of evidence suggests that linear imputation models do a reasonably good job of imputing non-normal variables (Schafer, 1997).

There are some important caveats, however. First, if a dichotomous variable has an extreme split (e.g., 3 percent ones and 97 percent zeros), a linear model may not give satisfactory imputations. Second, recent analytic and simulation results (Horton, Lipsitz, & Parzen, 2003; Allison, 2006) strongly indicate that rounding imputed values for dichotomous variables to 0 or 1 makes things worse rather than better, at least in terms of the quality of parameter estimates.[1] Similarly, transforming continuous but skewed variables to achieve approximate normality before imputation and then reversing the transformation after imputation often degrades parameter estimates, because the transformations make the imputation model inconsistent with the analysis model (von Hippel, 2009), an issue to be discussed below. Imposing upper and lower bounds on imputed values can also lead to bias because it inappropriately reduces variances.

In short, linear imputation models usually do a satisfactory job with non-normal variables, but the best practice is to leave the imputed values as they are, even if those values are unlike the real values in some respects. Sometimes, however, a linear imputation model is just not satisfactory (e.g., when imputing the dependent variable in a logistic regression). Then it is better to use an imputation method specifically designed for a particular kind of variable. We consider a few such models below.

20.5.4. How Many Data Sets?

As already noted, the MI procedure in SAS produces five imputed data sets by default, although it is easy to request more. Five is usually enough for the parameter estimates, but good estimates of standard errors, confidence intervals, and *p*-values often require more. More are always better, but how many are enough? The estimated degrees of freedom for each parameter, reported by most MI software, is a useful diagnostic. The *df* is used with the *t* distribution to calculate *p*-values and confidence intervals, and it increases as a linear function of the number of data sets. When I do MI, I like every *df* to be at least 100. At that point, the *t* distribution approaches the normal distribution, and little is to be gained from additional data sets. For our divorce example, the lowest *df* was 179, suggesting no need for additional data sets.

20.5.5. Multivariate Inference

A general principle of MI is that any population quantity can be estimated by simply averaging its estimates over the repeated data sets. Besides regression coefficients, this includes summary statistics like R^2 and root mean squared error, although MI software often does not report these. It is never correct to average test statistics, like *t*, *F*, or chi-square statistics, however. Special methods are required to combine such statistics across multiple data sets.

Multivariate inference, that is, testing hypotheses about more than one coefficient at a time, often requires such methods. For example, we frequently need to test

[1] This is contrary to my recommendation in Allison (2001).

whether two coefficients are equal, or whether several coefficients are all equal to zero. Three well-known approaches for constructing test statistics for such hypotheses are likelihood ratio tests, Wald tests, and a method for combining chi-squares. Details can be found in Schafer (1997), Allison (2001), or Little and Rubin (2002). Likelihood ratio tests are considered the most accurate, but they are also the most difficult to calculate.

20.5.6. Congeniality of the Imputation Model and the Analysis Model

For MI to perform optimally, the model used to impute the data must be "congenial" in some sense with the model intended for analysis (Rubin, 1987; Meng, 1994). The models need not be identical, but the imputation model must generate imputations that reproduce the major features of the data that are the focus of the analysis. That is the main reason I recommend that the imputation model include all variables in the model of interest.

Using an imputation model that is less restrictive than the analysis model creates no problems. As we have seen, it is often good for the imputation model to include auxiliary variables that are not in the analysis model. But trouble can arise if the imputation model is *more* restrictive than the analysis model. For example, if an analysis model contains non-linearities and/or interactions, the imputation model should also include them. With interactions, for instance, this means creating the product variables *before* doing the imputation, and then imputing these along with the original variables. Some MI software constrains imputed values of product variables to equal the product of the imputed values of the original variables, but von Hippel (2009) shows that this leads to biased estimates.

Ideally, multiple imputation should be tailored to the model of interest. An important implication of these principles is that a single set of imputed data sets may not be suitable for all models that users want to estimate.

20.5.7. Sequential Generalized Regression

Many different methods for producing multiple imputed data sets are consistent with the general principles described here. They differ mainly in the distributional models they assume for the data, and in the iterative algorithms they use to produce random draws from the posterior distribution of the missing data.

Sequential generalized regression (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001), also known as multiple imputation by chained equations (MICE) (Brand, 1999; Van Buuren & Oudshoorn, 2000) or fully conditional imputation, has recently become very popular. It is attractive because of its ability to impute both quantitative and categorical variables appropriately. It specifies a regression equation for imputing each variable with missing data, usually linear regression for quantitative variables, and logistic regression (binary, ordinal, or unordered multinomial) for categorical ones. Under logistic imputation, imputed values for categorical variables will also be categorical. Some software can also impute count variables by Poisson regression.

Imputation proceeds sequentially, usually starting from the variable with the least missing data and progressing to the variable with the most missing data. At each step, random draws are made from both the posterior distribution of the parameters and the posterior distribution of the missing values. Imputed values at one step are used as predictors in the imputation equations at subsequent steps (something that never happens in MCMC algorithms). Once all missing values have been imputed, several iterations of the process are repeated before selecting a completed data set.

Add-on software for sequential generalized regression is available for several statistical packages, including SAS, Stata, and Splus. I illustrate the method for our divorce example using the *ice* command in Stata (Royston, 1994; Carlin, Galati, & Royston, 2008). As with the SAS MI procedure, it specifies the set of variables to be used in the imputation, and a file name (‘‘coupleimp’’) for saving the imputed data sets:

```
ice div meduc feduc fgenid mprvdv fprvdv clt6 fprblow mprblow
  fbod mbod fcath mcath frelg mrelg, saving(coupleimp) m(5)
```

The ‘‘m(5)’’ option requests five data sets — the default is just one. As with SAS, these data sets are stacked one on top of another in the single Stata data set ‘‘coupleimp’’. A new variable ‘‘_mj’’ has values of 1 through 5 to distinguish the different data sets.

By default, *ice* imputes binary variables by logistic regression. Variables with more than five distinct values are imputed by linear regression. Thus, in this example, meduc, feduc, fgenid, fbod, mbod, frelg, and mrelg are all imputed by linear regression. The remaining six variables are imputed by logistic regression.

In Stata, unlike SAS, one command (with the *mim* prefix) performs the analysis on each data set, and also combines the results into a single set of estimates and associated statistics:

```
mim: logit div meduc feduc fgenid mprvdv fprvdv clt6 fprblow
  mprblow fbod mbod fcath mcath frelg mrelg
```

Results in the third panel of Table 20.2 are very similar to those obtained with the MCMC method using linear regression for all imputations. In fact, the results differ no more than one would expect from two different runs of multiple imputation using exactly the same method.

Though attractive, sequential generalized regression has two disadvantages compared with the linear MCMC method. First, it is much slower, computationally. Second, no theory justifies it. If all assumptions are met, the MCMC methods discussed earlier are guaranteed to converge to the correct posterior distribution. Sequential generalized regression carries no such guarantee, although simulation results by Van Buuren, Brand, Groothuis-Oudshoorn, and Rubin (2006) are very encouraging.

20.6. Maximum Likelihood

As methods for handling missing data, maximum likelihood and multiple imputation are close competitors. Under identical assumptions, both methods produce consistent, asymptotically efficient, and asymptotically normal estimates. Nevertheless, I prefer ML whenever it can be implemented for several reasons. First, ML produces a deterministic result while MI gives a different result every time it is used, because of its random draws from posterior distributions. Second, MI is often vulnerable to bias introduced by lack of congeniality between the imputation model and the analysis model. No such conflict is possible with ML because its estimates are based on a single, comprehensive model. Third, ML is generally a much ‘‘cleaner’’ method, requiring many fewer decisions about implementation.

The downside of ML is that it typically requires specialized, standalone software that is only available for a limited set of models. Both the number of software packages and the range of available models have greatly expanded recently, however. Several packages can do ML estimation with missing data for almost any linear model. And one of them, Mplus, can also estimate logistic regression models and Cox regression models with data missing both on the response variable and on predictors.

In general, the first step in ML estimation is to construct the likelihood function. Suppose that we have n independent observations ($i\ =1,\ldots,\ n$) on k variables ($y_{i1},\ y_{i2},\ldots,\ y_{ik}$) and no missing data. The likelihood function is

$$L = \prod_{i=1}^{n} f_i(y_{i1}, y_{i2}, \ldots, y_{ik}; \theta)$$

where $f_i(.)$ is the joint probability (or probability density) function for observation i, and θ is a set of parameters to be estimated. To get the ML estimates, we find the values of θ that make L as large as possible. Many methods can accomplish this, any one of which should produce the right result.

Now suppose that for a particular observation i, the first two variables, y_1 and y_2, have missing data that satisfy the ignorability assumption. The joint probability for that observation is just the probability of observing the remaining variables, y_{i3} through y_{ik}. If y_1 and y_2 are discrete, this is the above joint probability summed over all possible values of the two variables with missing data:

$$f_i^*(y_{i3}, \ldots, y_{ik}; \theta) = \sum_{y_1} \sum_{y_2} f_i(y_{i1}, \ldots, y_{ik}; \theta)$$

If the missing variables are continuous, we use integrals in place of summations:

$$f_i^*(y_{i3}, \ldots, y_{ik}; \theta) = \int_{y_1} \int_{y_2} f_i(y_{i1}, y_{i2}, \ldots y_{ik}; \theta) \mathrm{d}y_2 \mathrm{d}y_1$$

Essentially, then, for each observation's term in the likelihood function, we sum or integrate over the variables that have missing data, obtaining the marginal probability of observing those variables that have actually been observed.

As usual, the overall likelihood is just the product of the likelihoods for all the observations. For example, if there are *m* observations with complete data and *n–m* observations with data missing on y_1 and y_2, the likelihood function for the full data set becomes

$$L = \prod_{i=1}^{m} f_i(y_{i1}, y_{i2}, \ldots, y_{ik}; \theta) \prod_{m+1}^{n} f_i^*(y_{i3}, \ldots, y_{ik}; \theta)$$

where the observations are ordered such that the first *m* have no missing data and the last *n–m* have missing data.[2] This likelihood can then be maximized to get ML estimates of θ, again in several different ways.

20.6.1. EM Algorithm

One popular method for maximizing the likelihood when data are missing is the EM algorithm (Dempster, Laird, & Rubin, 1977). This iterative algorithm consists of two steps:

1. In the E (expectation) step, one finds the expected value of the log-likelihood, where the expectation is taken over the variables with missing data, based on the current values of the parameters.
2. In the M (maximization) step, the expected log-likelihood is maximized to produce new values of the parameters.

These steps are repeated, usually many times, until the algorithm converges, i.e., until the parameter estimates do not change from one iteration to the next.

Most comprehensive software packages have procedures that implement the EM algorithm. However, almost all assume multivariate normality, which implies that the relevant parameters are the means, variances, and covariances for all the variables. Under multivariate normality, the EM algorithm reduces to a kind of iterated linear regression imputation.

The main products of the EM algorithm are the ML estimates of the means, variances and covariances. These can then be used with linear modeling software to get estimates of regression coefficients and other parameters, which are true ML estimates of the relevant parameters. However, this two-step approach will generally produce incorrect standard errors and test statistics. As with pairwise deletion, this is because standard error estimates require that a sample size be specified, and no single value yields correct standard errors for all parameters.

[2] Additional missing data patterns would require separate, similarly constructed terms in the likelihood.

20.6.2. Direct Maximum Likelihood

A better method is direct ML, also known as raw ML (because it requires raw data rather than a covariance matrix) or full-information ML (Arbuckle, 1996; Allison, 2003). This method directly specifies the likelihood for the model to be estimated, and then maximizes it by conventional numerical methods (like Newton–Raphson) that produce standard errors as a by-product. This approach has been implemented for a wide class of linear structural equation models (including ordinary linear regression) in several packages: LISREL (www.ssicentral.com), EQS (www.mvsoft.com), Amos (www.spss.com/amos), MX (www.vcu.edu/mx), and Mplus (www.statmodel.com). For some of these packages (e.g., Amos and Mplus), direct ML is the default method for handling missing data, and no special program instructions are necessary.

The special appeal of Mplus is its capacity to do direct ML for logistic regression and Cox regression. We illustrate its use for our divorce example. For logistic regression, listwise deletion is the default in Mplus. Several special instructions are necessary to do direct ML. Here is the complete program code for doing the analysis:

```
data: file = c: \couple.txt;
variable:
  names = div meduc feduc fgenid mprvdv fprvdv
    clt6 fprblow mprblow fbod mbod
    fcath mcath frelg mrelg;
  missing = . ; categorical = div;
analysis: estimator = ml; integration = montecarlo;
model:
  div on meduc feduc fgenid mprvdv fprvdv
    clt6 fprblow mprblow fbod mbod
    fcath mcath frelg mrelg;
  meduc feduc fgenid mprvdv fprvdv
    clt6 fprblow mprblow fbod mbod
    fcath mcath frelg mrelg;
```

After reading in the data as a text file, we assign a missing data code (there is no default) and specify that the dependent variable "div" is a categorical variable. Specifying "estimator = ml" implies that "div" will modeled by a logistic regression. The "integration = montecarlo" option is necessary to integrate the likelihood function over the missing values. The model consists of two parts, separated by a semicolon. The first specifies the logistic regression. The second part specifies that the predictor variables have a multivariate normal distribution. While this may seem undesirable since many of the variables are dichotomous, it is necessary to specify some kind of joint distribution for the predictors. The multivariate normal is the simplest and most computationally efficient specification. For each individual, the likelihood is the product of the conditional distribution of "div" multiplied by the joint distribution of the predictors, with numerical integration over the variables that have missing data. Results are shown in the fourth panel of Table 20.2. The ML estimates (both coefficients and standard errors) are very similar to those for MI.

If the multivariate normal assumption for the predictors seems too strong, one can specify a more complex model that treats the dichotomous predictors as categorical. Here is the Mplus code for the divorce example:

```
data: file = c: \couple.txt;
variable:
  names = div meduc feduc fgenid mprvdv fprvdv
    clt6 fprblow mprblow fbod mbod
    fcath mcath frelg mrelg;
  missing = . ; categorical = div mprvdv fprvdv clt6
    fprblow mprblow fcath mcath;
analysis: estimator = ml; integration = montecarlo;
model:
  div on meduc feduc fgenid mprvdv fprvdv
       clt6 fprblow mprblow fbod mbod
       fcath mcath frelg mrelg;
  mprvdv on fprvdv clt6 fprblow mprblow fcath mcath
    meduc feduc fgenid frelg mrelg fbod mbod;
  fprvdv on clt6 fprblow mprblow fcath mcath
    meduc feduc fgenid frelg mrelg fbod mbod;
  clt6 on fprblow mprblow fcath mcath
    meduc feduc fgenid frelg mrelg fbod mbod;
  fprblow on mprblow fcath mcath
    meduc feduc fgenid frelg mrelg fbod mbod;
  mprblow on fcath mcath
    meduc feduc fgenid frelg mrelg fbod mbod;
  fcath on mcath
    meduc feduc fgenid frelg mrelg fbod mbod;
  mcath on meduc feduc fgenid frelg mrelg fbod mbod;
  meduc feduc fgenid frelg mrelg fbod mbod;
```

This program declares all the dichotomous variables to be categorical. In the "model" command, the logistic regression for "div" is followed by a series of recursive logistic regression equations, expressing each dichotomous predictor variable as a function of "prior" variables. The ordering of these variables is completely arbitrary. Though the ordering could in principle make a difference, different orderings produced exactly the same results for this example. The last line of the program specifies that the quantitative variables have a multivariate normal distribution. Results in the fifth panel of Table 20.2 (ML 2) are extremely close to those obtained with the much simpler multivariate normality specification (ML 1). So imposing that assumption made virtually no difference in this case.

20.7. Longitudinal Data

Longitudinal studies are particularly prone to missing data problems because it is difficult to follow individuals over substantial periods of time (see Stafford, this volume). Some people stop participating, some cannot be located, others may be away at the time of re-contact. Either MI or ML can handle missing data in longitudinal studies quite well. These methods usually can be implemented for longitudinal data in a fairly straightforward manner.

Whatever method is used, it is important to use *all* available information over time in order to minimize bias and standard errors. For example, suppose that one wishes to estimate a random-effects regression model using panel data for 1000 people and five time points per person, but some predictor variables have missing data. Most random-effects software requires a separate observational record for each person and point in time (the so-called ''long'' form), with a common ID number for all observations for the same person. One *should not* do multiple imputation on those 5000 records. That would impute missing values using only information obtained at the same point in time.

A much better method is to restructure the data so that there is one record per person (the ''wide'' form): a variable like income measured at five points in time would be represented by five different variables. Then multiple imputation with a variable list including all the variables in the model at all five time points would impute values for any variable with missing data using all the other variables at all time points, including (especially) the same variable measured at other time points. Once these missing data are imputed, the data set can be reshaped again for purposes of analysis into one with multiple observations per person.

Making imputations using data from all time points can substantially reduce the standard errors of the parameter estimates, and also reduce bias. If, for example, the dependent variable is a measure of depression, and people who are depressed at time 1 are more likely to drop out at time 2, then imputing the time 2 depression score using depression at time 1 can be very helpful in correcting for possible selection bias.

Using data at later time points to impute missing data at earlier ones may seem unsettling, since it seems to violate conventional notions of causal direction. But imputation has nothing to do with causality. It merely seeks to generate imputed values that are consistent with all the observed relationships among the variables.

Some may also be troubled by the fact that this method generates imputed values for all variables at all time points, even if a person dropped out after the first interview. Is that too much imputation? If a person died after the first interview, a reasonable case could be made for excluding records for times after death. But if someone simply dropped out of the study, imputing all subsequent missing data is better because selection bias may be substantially reduced. Remember that both MI and ML account completely for the fact that the some data are imputed when calculating standard errors, so the imputation of later waves does not artificially inflate the sample size.

Everything here about MI for longitudinal data also applies to ML, although it may be less obvious. Most structural equation modeling programs expect data sets with one record per person (i.e., in wide form). They will automatically use all data to "impute" missing values, and no special effort is needed. Some ML software for estimating random effects models (e.g., PROC MIXED in SAS) will appropriately handle missing data on the dependent variable even though the data include a record for each person and time point. But they simply delete cases with missing data on predictor variables from the analysis.

The considerations discussed in this section do not apply to our divorce example. Although the NSFH study was longitudinal, our analysis is essentially cross-sectional: the dependent variable (whether or not the couple divorced) is only measured once and the predictor variables are only measured in the baseline wave.

20.8. Methods For Nonignorable Missing Data

All methods described above assume that the missing data mechanism is ignorable and, hence, the data are missing at random. To reiterate, MAR means that the probability of missingness on any variable is unrelated to the value of that variable, once other observed variables are controlled. Unfortunately, this assumption is not testable.

Although MAR is a much weaker assumption than MCAR, there are often strong reasons to believe that is violated. If the variable of interest measures depression, for example, it seems plausible that depressed people would be less likely to agree to be interviewed, even after adjusting for prior and subsequent levels of depression. That could lead to biased estimates.

Both MI and ML methods exist for the not missing at random (NMAR) situation. If their assumptions are satisfied, their estimates have their usual desirable properties of consistency, asymptotic efficiency, and asymptotic normality. However, to accomplish this, they require a valid model for the missing data mechanism, and that is not easily obtained, at least not with any confidence. Many models for the missing data mechanism may be plausible, but their fit to the data cannot be compared, and results may depend greatly on which one is chosen.

Before examining some NMAR methods, recall from Section 20.3 that listwise deletion actually handles data that are not missing at random on *predictor* variables pretty well in any kind of regression analysis. Specifically, if the probability of missing data on a predictor is unrelated to the *dependent* variable, listwise deletion will not produce bias in coefficient estimates, though reduction in power due to the loss of cases may be substantial.

Methods for nonignorable missing data fall into two classes: selection models and pattern-mixture models (Little & Rubin, 2002). Some notation helps to explain the difference. Let $\mathbf{y}_i$ be a $k \times 1$ random vector containing all variables in the model of interest for individual i. Let $\mathbf{r}_i$ be a $k \times 1$ random vector of dummy (indicator) variables having values of 1 if the corresponding variable in $\mathbf{y}_i$ is observed and 0 if that variable is missing. Each individual's contribution to the likelihood function is then given by $f(\mathbf{y}_i, \mathbf{r}_i)$, the density for the joint distribution of $\mathbf{y}_i$ and $\mathbf{r}_i$.

Using the definition of conditional probability, we can factor this joint density in two different ways. For selection models, we have

$$f(\mathbf{y}_i, \mathbf{r}_i) = \Pr(\mathbf{r}_i|\mathbf{y}_i)f(\mathbf{y}_i),$$

where $f(\mathbf{y}_i)$ is the marginal density function for $\mathbf{y}_i$, and $\Pr(\mathbf{r}_i|\mathbf{y}_i)$ the conditional probability of $\mathbf{r}_i$ given $\mathbf{y}_i$. (This conditional probability is what we mean by the ''missing data mechanism'', which can be ignored when the data are missing at random.) Modeling then proceeds by specifying each of these two components in more detail. For example, $f(\mathbf{y}_i)$ could be specified as a multivariate normal distribution, and $\Pr(\mathbf{r}_i|\mathbf{y}_i)$ as a set of logistic regression models.

For pattern-mixture models, the factorization is

$$f(\mathbf{y}_i, \mathbf{r}_i) = f(\mathbf{y}_i|\mathbf{r}_i)\Pr(\mathbf{r}_i),$$

i.e., the conditional density of $\mathbf{y}_i$ given $\mathbf{r}_i$, times the probability of $\mathbf{r}_i$. Again, modeling proceeds by further specification of each of the two factors.

Although these two factorizations are both mathematically correct, pattern-mixture models are conceptually less appealing because they suggest that the distribution of $\mathbf{y}$ depends on whether we observe it or not. And even if one can estimate $f(\mathbf{y}|\mathbf{r})$ (which is not always possible), it is usually necessary to sum those distributions over the patterns in $\mathbf{r}$ to get the parameter estimates that are actually desired. Hence, I will not consider pattern-mixture models further here. See Little (1993) for details on formulating and estimating such models.

The best-known selection model is undoubtedly Heckman's (1979) model for selection bias. This model describes $f(y_i)$ by a conventional linear regression,

$$y_i = \boldsymbol{\beta}\mathbf{x}_i + \varepsilon_i,$$

where $\mathbf{x}_i$ is a vector of predictor variables. The random disturbance ε_i is assumed to have a normal distribution with a mean of 0, constant variance σ^2, and a covariance of 0 across observations. This implies that $f(y_i)$ is a normal density function with mean $\boldsymbol{\beta}\mathbf{x}_i$ and variance σ^2. With no missing data, the coefficient vector $\boldsymbol{\beta}$ could be optimally estimated by ordinary least squares.

Now suppose that some data are missing on y_i, and we want to allow the probability that y is missing to depend on y itself, even after adjusting for $\mathbf{x}$. Let $R = 1$ if y is observed and 0 if y is missing. A probit model for R states that:

$$\Pr(R_i = 1|y_i, \mathbf{x}_i) = \Phi(\delta y_i + \boldsymbol{\alpha}\mathbf{x}_i)$$

where $\Phi(.)$ is the cumulative distribution function for a standard normal variable, δ is the coefficient for y, and $\boldsymbol{\alpha}$ is a vector of coefficients for $\mathbf{x}$. If $\delta = 0$, the data are MAR. If both δ and $\boldsymbol{\alpha}$ are 0, the data are MCAR.

Clearly, the probit equation cannot be estimated by itself because all ys are missing when $R = 0$. Remarkably, however, all the parameters are identified when the regression equation and the probit equation are combined in the likelihood function. The likelihood can be easily maximized using conventional numerical routines available in many statistics packages.

Unfortunately, Little and Rubin (2002) have shown that the Heckman model is exquisitely sensitive to the normality assumption on ε in the regression equation. If the ε distribution is skewed, for example, the Heckman method can easily yield more biased estimates than those obtained with listwise deletion. This is typical of selection models: because they are identified only by choosing specific functional forms and distributional shapes, slight variations on those choices can have major consequences.

Little and Rubin strongly advise that a sensitivity analysis should accompany any use of NMAR models. That is, one should try out a variety of different models to see how stable the results are. That good advice is not so easy to implement, given the limited availability of software for NMAR models.

Daniels and Hogan (2008) argue that pattern-mixture models are much more amenable to sensitivity analysis than selection models, and develop such models in some detail. But their approach is strictly Bayesian, requiring informative prior distributions and the use of WinBugs software (www.mrc-bsu.cam.ac.uk/bugs). Moreover, their models are quite complex, even for very simple missing data patterns.

The bottom line is that although NMAR methods are available, they should be used with a great deal of caution. Proper use of these methods requires considerable knowledge and understanding of the missing data process, and a substantial commitment to try out different plausible models.

20.9. Summary

Conventional methods for handling missing data leave much to be desired. They typically yield biased parameter estimates or standard error estimates that are too low. They often require unrealistically strong assumptions about the missing data mechanism. In my opinion, the best of them is still listwise deletion. It is unbiased if data are missing completely at random. It yields unbiased estimates of regression coefficients when missing data are for predictor variables in a regression model, even when the data are not missing at random. Most importantly, it gives honest estimates of standard errors.

Despite these good qualities, listwise deletion can needlessly discard a lot of data, often far more than researchers are willing to tolerate. To avoid this, use multiple imputation or maximum likelihood. If assumptions are met, these two methods yield approximately unbiased estimates, along with standard error estimates that accurately reflect the amount of information in the data. Standard implementations of these methods require the missing at random assumption. Although though this assumption is strong, it is much weaker than the MCAR assumption required for most conventional methods.

Although the two methods have similar assumptions and properties, I prefer maximum likelihood whenever it is available. It produces deterministic results, and entails no potential conflict between an imputation model and an analysis model. The downside is that ML typically requires stand-alone software, and it is not available for many models. MI, on the other hand, is available in most major statistical packages and can be used for virtually any kind of model.

For our empirical example, maximum likelihood and multiple imputation produced very similar results for both coefficient estimates and their standard errors. Furthermore, for either method, it made little difference whether the dichotomous variables with missing data were modeled by linear regressions or logistic regressions. These similarities are reassuring and fairly typical.

References

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Allison, P. D. (2003). Missing data techniques for structural equation models. *Journal of Abnormal Psychology*, *112*, 545–557.

Allison, P. D. (2006). Multiple imputation of categorical variables under the multivariate normal model. Paper presented at the annual meeting of the American Sociological Association, Montreal Convention Center, Montreal, Quebec, Canada, Aug. 11, 2006. Available at http://www.allacademic.com/meta/p_mla_apa_research_citation/1/0/2/5/4/p102543_index.html

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In: G. A. Marcoulides & R. E. Schumacker (Eds), *Advanced structural equation modeling: Issues and techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.

Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets, dissertation*. Rotterdam: Erasmus University.

Carlin, J. B., Galati, J. C., & Royston, P. (2008). A new framework for managing and analyzing multiply imputed data in Stata. *Stata Journal*, *8*, 49–67.

Cohen, J., & Cohen, P. (1985). *Applied multiple regression and correlation analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies*. Boca Raton, FL: Chapman & Hall/CRC.

Davis, J. A., Smith, T. W., & Marsden, P. V. (2007). General social surveys, 1972–2006 [Cumulative file] [Computer file]. ICPSR04697-v2. Chicago, IL: National Opinion Research Center [producer]. Storrs, CT: Roper Center for Public Opinion Research, University of Connecticut/Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors], 2007-09-10. doi:10.3886/ICPSR04697.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. *Journal of the American Statistical Association*, *59*, 834–844.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*, 197–218.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statsitical Society, Series B*, *30*, 67–82.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161.

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, *57*, 229–232.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, *91*, 222–230.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, *87*, 1227–1237.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125–154.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558.

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

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*, 122–129.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*, 106–129.

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

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semi-parametric nonresponse models (with comments). *Journal of the American Statistical Association*, *94*, 1096–1146.

Sweet, J. A., & Bumpass, L. L. (2002). The national survey of families and households-waves 1, 2, and 3: Data description and documentation. Center for demography and ecology, University of Wisconsin-Madison. Available at http://www.ssc.wisc.edu/nsfh/home.htm

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1046–1064.

Van Buuren, S., & Oudshoorn, C. G. M. (2000). Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual, Report PG/VGZ/00.038, Leiden: TNO Preventie en Gezondheid.

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *2009*, 265–291.