4

Missing Data

Paul D. Allison

INTRODUCTION

Missing data are ubiquitous in psychological research. By missing data, I mean data that are missing for some (but not all) variables and for some (but not all) cases. If data are missing on a variable for *all* cases, then that variable is said to be latent or unobserved. On the other hand, if data are missing on *all* variables for some cases, we have what is known as *unit* non-response, as opposed to *item* non-response which is another name for the subject of this chapter. I will not deal with methods for latent variables or unit non-response here, although some of the methods we will consider can be adapted to those situations.

Why are missing data a problem? Because conventional statistical methods and software presume that all variables in a specified model are measured for all cases. The default method for virtually all statistical software is simply to delete cases with any missing data on the variables of interest, a method known as listwise deletion or complete case analysis. The most obvious drawback of listwise deletion is that it often deletes a large fraction of the sample, leading to a severe loss of statistical power. Researchers are understandably reluctant to discard data that they have spent a great deal time, money and effort in collecting. And so, many methods for 'salvaging' the cases with missing data have become popular.

For a very long time, however, missing data could be described as the 'dirty little secret' of statistics. Although nearly everyone had missing data and needed to deal with the problem in some way, there was almost nothing in the textbook literature to provide either theoretical or practical guidance. Although the reasons for this reticence are unclear, I suspect it was because none of the popular methods had any solid mathematical foundation. As it turns out, all of the most commonly used methods for handling missing data have serious deficiencies.

Fortunately, the situation has changed dramatically in recent years. There are now two broad approaches to missing data that have excellent statistical properties if the specified assumptions are met: maximum likelihood and multiple imputation. Both of these methods have been around in one form or another for at least 30 years, but it is only in the last decade that they have become fully developed and incorporated into widely-available and easily-used software. A third method, inverse probability weighting (Robins and Rotnitzky, 1995; Robins et al.,

1995; Scharfstein et al., 1999), also shows much promise for handling missing data but has not yet reached the maturity of the other two methods. In this chapter, I review the strengths and weaknesses of conventional missing data methods but focus the bulk of my attention on maximum likelihood and multiple imputation.

MISSING COMPLETELY AT RANDOM

Before beginning an examination of specific methods, it is essential to spend some time discussing possible assumptions. No method for handling missing data can be expected to perform well unless there are some restrictions on how the data came to be missing. Unfortunately, the assumptions that are necessary to justify a missing data method are typically rather strong and often untestable. The strongest assumption that is commonly made is that the data are *missing completely at random* (MCAR). This assumption is most easily explained for the situation in which there is only a single variable with missing data, which we will denote by Z. Suppose we have another set of variables (represented by the vector $\mathbf{X}$) which is always observed. Let R_Z be an indicator (dummy) variable having a value of 1 if Z is missing and 0 if Z is observed. The MCAR assumption can then be expressed by the statement:

$$\Pr(R_Z = 1|X, Z) = \Pr(R_Z = 1)$$

That is, the probability that Z is missing depends neither on the observed variables X nor on the possibly missing values of Z itself.

A common question is: what variables have to be in X in order for the MCAR assumption to be satisfied? All variables in the data set? All possible variables, whether in the data set or not? The answer is: only the variables in the model to be estimated. If you are estimating a multiple regression and Z is one of the predictors, then the vector X must include all the other variables in the model. But missingness on Z could depend on some other variable (whether in the data set or not), and it would not be a violation of MCAR. On the other hand, if you are merely estimating the mean of Z, then all that is necessary for MCAR is $\Pr(R_Z = 1|Z) = \Pr(R_Z = 1)$. That is, missingness on Z does not depend on Z itself.

When more than one variable in the model of interest has missing data, the statement of MCAR is a bit more technical and will not be given here (see Rubin, 1976). But the basic idea is the same: the probability that any variable is missing cannot depend on any other variable in the model of interest, or on the potentially missing values themselves. It is important to note, however, that the probability that one variable is missing can depend on whether or not another variable is missing, without violating MCAR. In the extreme, two or more variables may always be missing together or observed together. This is actually quite common. It typically occurs when data sets are pieced together from multiple sources, for example, administrative records and personal interviews. If the subject declines to be interviewed, all the responses in that interview will be jointly missing.

For most data sets, the MCAR assumption is unlikely to be precisely satisfied. One situation in which the assumption *is* likely to be satisfied is when data are *missing by design* (Graham et al., 1996). For example, a researcher may decide that a brain scan is just too costly to administer to everyone in her study. Instead, she does the scan for only a 25% random subsample. For the remaining 75%, the brain-scan data are MCAR.

Can the MCAR assumption be tested? Well, it is easy to test whether missingness on Z depends on X. The simplest approach is to test for differences in means of the Xvariables between those who responded to Z and those who did not, a strategy that has been popular for years. A more comprehensive approach is to do a logistic regression of R_Z on X. Significant coefficients, either singly or jointly, would indicate a violation of MCAR. On the other hand, it is not possible to test whether missingness on Z depends on Z itself (conditional on X). That would require knowledge of the missing values.

MISSING AT RANDOM

A much weaker (but still strong) assumption is that the data are *missing at random* (MAR). Again, let us consider the special case in which only a single variable Z has missing data, and there is a vector of variables X that is always observed. The MAR assumption may be stated as:

$$\Pr(R_Z = 1|X, Z) = \Pr(R_Z = 1|X)$$

This equation says that missingness on Z may depend on X, but it does not depend on Z itself (after adjusting for X). For example, suppose that missingness on a response variable depends on whether a subject is assigned to the treatment or the control group, with a higher fraction missing in the treatment group. But within each group, suppose that missingness does not depend on the value of the response variable. Then, the MAR assumption is satisfied. Note that MCAR is a special case of MAR. That is, if the data are MCAR, they are also MAR.

As with MCAR, the extension to more than one variable with missing data requires more technical care in stating the assumption (Rubin, 1976), but the basic idea is the same: the probability that a variable is missing may depend on anything that is observed; it just cannot depend on any of the unobserved values of the variables with missing data (after adjusting for observed values). Nevertheless, missingness on one variable is allowed to depend on missingness on other variables.

Unfortunately, the MAR assumption is *not* testable. You may have reasons to suspect that the probability of missingness depends on the values that are missing, for example, people with high incomes may be less likely to report their incomes. But nothing in the data will tell you whether this is the case or not. Fortunately, there is a way to make the assumption more plausible. The MAR assumption says that missingness on Z does not depend on Z, *after adjusting for the variables in* X. And like MCAR, the set of variables in X depends on the model to be estimated. If you put many variables in X, especially those that are highly correlated with Z, you may be able reduce or eliminate the residual dependence of missingness on Z itself. In the case of income, for example, putting such variables as age, sex, occupation and education into the X vector can make the MAR assumption much more reasonable. Later on, we shall discuss strategies for doing this.

The missing-data mechanism (the process generating the missingness) is said to be *ignorable* if the data are MAR *and* an additional, somewhat technical, condition is satisfied. Specifically, the parameters governing the missing-data mechanism must be distinct from the parameters in the model to be estimated. Since this condition is unlikely to be violated in real-world situations it is commonplace to use the terms MAR and ignorability interchangeably. As the name suggests, if the missing-data mechanism is ignorable, then it is possible to get valid, optimal estimates of parameters without directly modeling the missing-data mechanism.

NOT MISSING AT RANDOM

If the MAR assumption is violated, the data are said to be not missing at random (NMAR). In that case, the missing-data mechanism is not ignorable, and valid estimation requires that the missing-data mechanism be modeled as part of the estimation process. A well-known method for handling one kind of NMAR is Heckman's (1979) method for selection bias. In Heckman's method, the goal is to estimate a linear model with NMAR missing data on the dependent variable Y. The missing-data mechanism may be represented by a probit model in which missingness on Y depends on both Y and X. Using maximum likelihood, the linear model and the probit model are estimated simultaneously to produce consistent and efficient estimates of the coefficients.

As there are many situations in which the MAR assumption is implausible it is tempting to turn to missing-data methods that do not

require this assumption. Unfortunately, these methods are fraught with difficulty. Because every NMAR situation is different, the model for the missing-data mechanism must be carefully tailored to each situation. Furthermore, there is no information in the data that would help you choose an appropriate model, and no statistic that will tell you how well a chosen model fits the data. Worse still, the results are often exquisitely sensitive to the choice of the model (Little and Rubin, 2002).

It is no accident, then, that most commercial software for handling missing data is based on the assumption of ignorability. If you decide to go the NMAR route you should do so with great caution and care. It is probably a good idea to enlist the help and advice of someone who has real expertise in this area. It is also recommended that you try different models for the missing-data mechanism to get an idea of how sensitive the results are to model choice. The remainder of this chapter will assume ignorability, although it is important to keep in mind that both maximum likelihood and multiple imputation can produce valid estimates in the NMAR case if you have a correctly specified model for the missing-data mechanism.

CONVENTIONAL METHODS

This section is a brief review of conventional methods for handling missing data, with an emphasis on what is good and bad about each method. To do that, we need some criteria for evaluating a missing-data method. There is general agreement that a good method should do the following:

1. Minimize bias. Although it is well-known that missing data can introduce bias into parameter estimates, a good method should make that bias as small as possible.
2. Maximize the use of available information. We want to avoid discarding any data, and we want to use the available data to produce parameter estimates that are efficient (i.e., have minimum-sampling variability).
3. Yield good estimates of uncertainty. We want accurate estimates of standard errors, confidence intervals and *p*-values.

In addition, it would be nice if the missing-data method could accomplish these goals without making unnecessarily restrictive assumptions about the missing-data mechanism. As we shall see, maximum likelihood and multiple imputation do very well at satisfying these criteria. But conventional methods are all deficient on one or more of these goals.

Listwise deletion

How does listwise deletion fare on these criteria? The short answer is good on 3 (above), terrible on 2 and so-so on 1. Let us first consider bias. If the data are MCAR, listwise deletion will not introduce any bias into parameter estimates. We know that because, under MCAR, the subsample of cases with complete data is equivalent to a simple random sample from the original target sample. It is also well-known that simple random sampling does not cause bias. If the data are MAR but not MCAR, listwise deletion *may* introduce bias. Here is a simple example. Suppose the goal is to estimate mean income for some population. In the sample, 85% of women report their income but only 60% of men (a violation of MCAR), but within each gender missingness on income does not depend on income (MAR). Assuming that men, on average, make more than women, listwise deletion would produce a downwardly biased estimate of mean income for the whole population.

Somewhat surprisingly, listwise deletion is very robust to violations of MCAR (or even MAR) for predictor variables in a regression analysis. Specifically, so long as missingness on the predictors does not depend on the dependent variable, listwise deletion will yield approximately unbiased estimates of regression coefficients (Little, 1992). And this holds for virtually any kind of regression – linear, logistic, Poisson, Cox, etc.

The obvious downside of listwise deletion is that it often discards a great deal of potentially usable data. On the one hand, this loss of data leads to larger standard errors, wider confidence intervals, and a loss of power in testing hypotheses. On the other hand, the estimated standard errors produced by listwise deletion are usually accurate estimates of the true standard errors. In this sense, listwise deletion is an 'honest' method for handling missing data, unlike some other conventional methods.

Pairwise deletion

For linear models, a popular alternative to listwise deletion is pairwise deletion, also known as available case analysis. For many linear models (e.g., linear regression, factor analysis, structural equation models), the parameters of interest can be expressed as functions of the population means, variances and covariances (or, equivalently, correlations). In pairwise deletion, each of these 'moments' is estimated using all available data for each variable or each pair of variables. Then, these sample moments are substituted into the formulas for the population parameters. In this way, all data are used and nothing is discarded.

If the data are MCAR, pairwise deletion produces consistent (and, hence, approximately unbiased) estimates of the parameters (Glasser, 1964). Like listwise deletion, however, if the data are MAR but not MCAR, pairwise deletion may yield biased estimates. Intuitively, pairwise deletion ought to be more efficient than listwise deletion because more data are utilized in producing the estimates. This is usually the case, although simulation results suggest that in certain situations pairwise deletion may actually be less efficient than listwise.

Occasionally, pairwise deletion breaks down completely because the estimated correlation matrix is not a definite positive and cannot be inverted to calculate the parameters. The more common problem, however, is the difficulty in getting accurate estimates of the standard errors. That is because each covariance (or correlation) may be based on a different sample size, depending on the missing-data pattern. Although methods have been proposed for getting accurate standard error estimates (Van Praag et al., 1985), they are complex and have not been incorporated into any commercial software.

Dummy-variable adjustment

In their 1985 textbook, Cohen and Cohen popularized a method for dealing with missing data on predictors in a regression analysis. For each predictor with missing data, a dummy variable is created to indicate whether or not data are missing on that predictor. All such dummy variables are included as predictors in the regression. Cases with missing data on a predictor are coded as having some constant value (usually the mean for non-missing cases) on that predictor.

The rationale for this method is that it incorporates all the available information into the regression. Unfortunately, Jones (1996) proved that this method typically produces biased estimates of the regression coefficients, even if the data are MCAR. He also proved the same result for a closely-related method for categorical predictors whereby an extra category is created to hold the cases with missing data. Although these methods probably produce reasonably accurate standard error estimates, the bias makes them unacceptable.

Imputation

A wide variety of methods falls under the general heading of imputation. This class includes any method in which some guess or estimate is substituted for each missing value, after which the analysis is done using conventional software. A simple but popular approach is to substitute means for missing values, but this is well- known to produce biased estimates (Haitovsky, 1968). Imputations based on linear regression are much better, although still problematic. One problem, suffered by most conventional,

deterministic methods is that they produce biased estimates of some parameters. In particular, variances for the variables with missing data tend to be underestimated, and this bias is propagated to any parameters that depend on variances (e.g., regression coefficients).

Even more serious is the tendency for imputation to produce underestimates of standard errors, which leads in turn to inflated test statistics and p-values that are too low. That is because conventional software has no way of distinguishing real data from imputed data and cannot take into account the inherent uncertainty of the imputations. The larger the fraction of missing data, the more severe this problem will be. In this sense, all conventional imputation methods are 'dishonest' and should be viewed with some skepticism.

MAXIMUM LIKELIHOOD

Maximum likelihood has proven to be an excellent method for handling missing data in a wide variety of situations. If the assumptions are met, maximum likelihood for missing data produces estimates that have the desirable properties normally associated with maximum likelihood: consistency, asymptotic efficiency and asymptotic normality. Consistency implies that estimates will be approximately unbiased in large samples. Asymptotic efficiency means that the estimates are close to being fully efficient (i.e., having minimal standard errors). Asymptotic normality is important because it means we can use a normal approximation to calculate confidence intervals and p-values. Furthermore, maximum likelihood can produce accurate estimates of standard errors that fully account for the fact that some of the data are missing.

In sum, maximum likelihood satisfies all three criteria stated earlier for a good missing-data method. Even better is the fact that it can accomplish these goals under weaker assumptions than those required for many conventional methods. In particular, it does well when data are MAR but not MCAR. It also does well when the data are not NMAR – if one has a correct model for the missing-data mechanism.

Of course there are some downsides. Specialized software is typically required. Also required is a parametric model for the joint distribution of all the variables with missing data. Such a model is not always easy to devise, and results may be somewhat sensitive to model choice. Finally, the good properties of maximum likelihood estimates are all 'large sample' approximations, and those approximations may be poor in small samples.

Most software for maximum likelihood with missing data assumes ignorability (and, hence, MAR). Under that assumption, the method is fairly easy to describe. As usual, to do maximum likelihood we first need a likelihood function, which expresses the probability of the data as a function of the unknown parameters. Suppose we have two discrete variables X and Z, with a joint probability function denoted by $p(x, z|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of parameters. That is, $p(x, z|\boldsymbol{\theta})$ gives the probability that $X = x$ and $Z = z$. If there are no missing data and observations are independent, the likelihood function is given by:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i, z_i|\boldsymbol{\theta})$$

To get maximum likelihood estimates, we find the value of $\boldsymbol{\theta}$ that makes this function as large a possible.

Now suppose that data are MAR on Z for the first r cases, and MAR on X for the next s cases. Let:

$$g(x|\theta) = \sum_{z} p(x, z|\boldsymbol{\theta})$$

be the marginal distribution of X (summing over Z) and let:

$$h(z|\theta) = \sum_{x} p(x, z|\boldsymbol{\theta})$$

be the marginal distribution of Z (summing over X). The likelihood function is then:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{r} g(x_i|\boldsymbol{\theta}) \prod_{i=r+1}^{r+s} h(z_i|\boldsymbol{\theta}) \prod_{i=r+s+1}^{n} p(x_i, z_i|\boldsymbol{\theta})$$

That is, the likelihood function is factored into parts that corresponding to different missing-data patterns. For each pattern, the likelihood is found by summing the joint distribution over all possible values of the variable(s) with missing data. If the variables are continuous rather than discrete, the summation signs are replaced with integral signs. The extension to more than two variables is straightforward.

To implement maximum likelihood for missing data, one needs a model for the joint distribution of all the relevant variables and a numerical method for maximizing the likelihood. If all the variables are categorical, an appropriate model might be the unrestricted multinomial model, or a log-linear model that imposes some restrictions on the data. The latter is necessary when there are many variables with many categories. Otherwise, without restrictions (e.g., all three-way and higher interactions are 0), there would be too many parameters to estimate. An excellent freeware package for maximizing the likelihood for any log-linear model with missing data is LEM (available at http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html). LEM can also estimate logistic-regression models (in the special case when all predictors are discrete) and latent-class models.

When all variables are continuous, it is typical to assume a multivariate-normal model. This implies that each variable is normally distributed and can be expressed as a linear function of the other variables (or any subset of them), with errors that are homoscedastic and have a mean of 0. While this is a strong assumption, it is commonly used as the basis for multivariate analysis and linear-structural equation modeling.

Under the multivariate-normal model, the likelihood can be maximized using either the expectation-maximization (EM) algorithm or direct maximum likelihood. Direct maximum likelihood is strongly preferred because it gives accurate standard error estimates and is more appropriate for 'overidentified' models. However, because the EM method is readily available in many commercial software packages, it is worth taking a closer look at it.

EM is a numerical algorithm that can be used to maximize the likelihood under a wide variety of missing-data models (Dempster et al., 1977). It is an iterative algorithm that repeatedly cycles through two steps. In the expectation step, the expected value of the log-likelihood is taken over the variables with missing data, using the current values of the parameter estimates to compute the expectation. In the maximization step, the expected log-likelihood is maximized to get new values of the parameter estimates. These two steps are repeated over and over until convergence, i.e., until the parameter estimates do not change from one iteration to the next.

Under the multivariate-normal model, the parameters that are estimated by the EM algorithm are the means, variances and covariances. In this case, the algorithm reduces to something that can be described as iterated linear regression imputation. The steps are as follows:

1. Get starting values for the means, variances and covariances. These can be obtained by listwise or pairwise deletion.
2. For each missing-data pattern, construct regression equations for predicting the missing variables based on the observed variables. The regression parameters are calculated directly from the current estimates of the means, variances and covariances.
3. Use these regression equations to generate predicted values for all the variables and cases with missing data.
4. Using all the real and imputed data, recalculate the means, variances and covariances. For means, the standard formula works fine. For variances (and sometimes covariances) a correction factor

must be applied to compensate for the downward bias that results from using imputed values.
5. Go back to step 2 and repeat until convergence.

The principal output from this algorithm is the set of maximum likelihood estimates of the means, variances and covariances. Although imputed values are generated as part of the estimation process, it is not recommended that these values be used in any other analysis. They are not designed for that purpose, and they will yield biased estimates of many parameters. One drawback of the EM method is that, although it produces the correct parameter estimates, it does not produce standard error estimates.

EXAMPLE

To illustrate the EM algorithm (as well as other methods to be considered later), we will use a data set that has records for 581 children who were interviewed in 1990 as part of the National Longitudinal Survey of Youth (NLSY). A text file containing these data is available at http://www.ssc.upenn.edu/~allison. Here are the variables:

ANTI	antisocial behavior, measured with a scale ranging from 0 to 6.
SELF	self-esteem, measured with a scale ranging from 6 to 24.
POV	poverty status of family, coded 1 for in poverty, otherwise 0.
BLACK	1 if child is black, otherwise 0
HISPANIC	1 if child is Hispanic, otherwise 0
DIVORCE	1 if mother was divorced in 1990, otherwise 0
GENDER	1 if female, 0 if male
MOMWORK	1 if mother was employed in 1990, otherwise 0

BLACK and HISPANIC are two categories of a three-category variable, the reference category being non-Hispanic white. The ultimate goal is to estimate a linear-regression model with ANTI as the dependent variable and all the others as predictors.

The original data set had no missing data. I deliberately produced missing data on several of the variables, using a method that satisfied the MAR assumption. The variables with missing data and their percentage missing are: SELF (25%), POV (26%), BLACK and HISPANIC (19%) and MOMWORK (15%). Listwise deletion on this set of variables leaves only 225 cases, less than half the original sample.

Application of the multivariate normal EM algorithm to these data produced the maximum likelihood estimates of the means, variances and covariances in Table 4.1. It may be objected that method is not appropriate for the variables POV, BLACK and HISPANIC and MOMWORK because, as dummy variables, they cannot possibly have a normal distribution. Despite the apparent validity of this objection, a good deal of simulation evidence and practical experience suggests that method does a reasonably good job, even when the variables with missing data are dichotomous (Schafer, 1997). We will have more to say about this issue later on.

What can be done with these estimates? Because covariances are hard to interpret, it is usually desirable to convert the covariance matrix into a correlation matrix, something that is easily accomplished in many software packages. One of the nice things about maximum likelihood estimates is that any function of those estimates will also be a maximum likelihood estimate of the corresponding function in the population. Thus, if s_i is the maximum likelihood estimate of the standard deviation of x_i, and s_{ij} is the maximum likelihood estimate of the covariance between x_i and x_j, then $r = s_{ij}/(s_i s_j)$ is the maximum likelihood estimate of their correlation. Table 4.2 displays the maximum likelihood estimates of the correlations.

Next, we can use the EM estimates as input to a linear regression program to estimate the regression of ANTI on the other variables. Many regression programs allow a covariance or correlation matrix as input. If maximum likelihood estimates for the means and covariances are used as the input, the

Table 4.1 Expectation-maximization (EM) estimates of means and covariance matrix

	ANTI	*SELF*	*POV*	*BLACK*	*HISPANIC*	*DIVORCE*	*GENDER*	*MOMWORK*
Means	1.56799	20.1371	0.34142	0.35957	0.24208	0.23580	0.50430	0.33546
Covariance matrix								
ANTI	2.15932	−0.6402	0.15602	0.08158	−0.04847	0.01925	−0.12637	0.07415
SELF	−0.64015	9.7150	−0.10947	−0.09724	−0.13188	−0.14569	−0.03381	0.00750
POV	0.15602	−0.1095	0.22456	0.06044	−0.00061	0.05259	0.00770	0.05446
BLACK	0.08158	−0.0972	0.06044	0.22992	−0.08716	0.00354	0.00859	−0.01662
HISPANIC	−0.04847	−0.1319	−0.00061	−0.08716	0.18411	0.00734	−0.01500	0.01657
DIVORCE	0.01925	−0.1457	0.05259	0.00354	0.00734	0.18020	−0.00015	−0.00964
GENDER	−0.12637	−0.0338	0.00770	0.00859	−0.01500	−0.00015	0.24998	0.00407
MOMWORK	0.07415	0.0075	0.05446	−0.01662	0.01657	−0.00964	0.00407	0.22311

Table 4.2 Expectation-maximization (EM) estimates of correlation matrix

	ANTI	*SELF*	*POV*	*BLACK*	*HISPANIC*	*DIVORCE*	*GENDER*	*MOMWORK*
ANTI	1.0000	−0.1398	0.2241	0.1158	−0.0769	0.0309	−0.1720	0.1068
SELF	−0.1398	1.0000	−0.0741	−0.0651	−0.0986	−0.1101	−0.0217	0.0051
POV	0.2241	−0.0741	1.0000	0.2660	−0.0030	0.2614	0.0325	0.2433
BLACK	0.1158	−0.0651	0.2660	1.0000	−0.4236	0.0174	0.0358	−0.0734
HISPANIC	−0.0769	−0.0986	−0.0030	−0.4236	1.0000	0.0403	−0.0699	0.0817
DIVORCE	0.0309	−0.1101	0.2614	0.0174	0.0403	1.0000	−0.0007	−0.0481
GENDER	−0.1720	−0.0217	0.0325	0.0358	−0.0699	−0.0007	1.0000	0.0172
MOMWORK	0.1068	0.0051	0.2433	−0.0734	0.0817	−0.0481	0.0172	1.0000

Table 4.3 Regression of ANTI on other variables

	No missing data		*Listwise deletion*		*Maximum likelihood*			*Multiple imputation*	
Variable	*Coeff.*	*SE*	*Coeff.*	*SE*	*Coeff.*	*Two-step SE*	*Direct SE*	*Coeff.*	*SE*
SELF	**−0.054**	0.018	−0.045	0.031	**−0.066**	0.022	0.022	**−0.069**	0.021
POV	**0.565**	0.137	**0.727**	0.234	**0.635**	0.161	0.162	**0.625**	0.168
BLACK	0.090	0.140	0.053	0.247	0.071	0.164	0.160	0.073	0.155
HISPANIC	−0.346	0.153	−0.353	0.253	−0.336	0.176	0.170	−0.332	0.168
DIVORCE	0.068	0.144	0.085	0.243	−0.109	0.166	0.146	−0.107	0.147
GENDER	**−0.537**	0.117	−0.334	0.197	**−0.560**	0.135	0.117	**−0.556**	0.118
MOMWORK	0.184	0.129	0.259	0.216	0.215	0.150	0.142	0.242	0.143

Coefficients (Coeff.) in bold are statistically significant at the .01 level.
SE, standard error.

resulting regression coefficient estimates will also be maximum likelihood estimates. The problem with this two-step approach is that it is not easy to get accurate standard error estimates. As with pairwise deletion, one must specify a sample size to get conventional regression software to produce standard error estimates. But there is no single number that will yield the right standard errors for all the parameters. I generally get good results using the number of non-missing cases for the variable with the most missing data (in this example, 431 cases on POV). But this method may not work well under all conditions.

Results are shown in Table 4.3. The first set of regression estimates is based on the original data set with no missing data. Three variables have p-values below .01: SELF, POV and GENDER. Higher levels of antisocial behavior are associated with lower levels of self-esteem, being in poverty and being male. The negative coefficient for Hispanic is also marginally significant. The next set of estimates was obtained with listwise deletion. Although the coefficients are reasonably close to those in the original data set, the standard errors are much larger, reflecting the fact that more than half the cases are lost. As a result,

only the coefficient for POV is statistically significant.

Maximum likelihood estimates are shown in the third panel of Table 4.3. The coefficients are generally closer to the original values than those from listwise deletion. More importantly, the estimated standard errors (using 431 as the sample size in the two-step method) are much lower than those from listwise deletion, with the result that POV, SELF and GENDER all have *p*-values below .01. The standard errors are still larger than those from the original data set, but that is to be expected because a substantial fraction of the data is now missing.

DIRECT MAXIMUM LIKELIHOOD

As noted, the problem with the two-step method is that we do not get dependable standard error estimates. This problem can be solved by using direct maximum likelihood, also known as 'raw' maximum likelihood (because one must use the raw data as input rather than a covariance matrix) or 'full information' maximum likelihood (Arbuckle, 1996; Allison, 2003). In this approach, the linear model of interest is specified, and the likelihood function is directly maximized with respect to the parameters of the model. Standard errors may be calculated by conventional maximum likelihood methods (such as computing the negative inverse of the information matrix). The presumption is still that the data follow a multivariate normal distribution, but the means and covariance matrix are expressed as functions of the parameters in the specified linear model.

Direct maximum likelihood is now widely available in most stand-alone programs for estimating linear-structural equation models, including LISREL, AMOS, EQS, M-PLUS and MX. For the NLSY data, the maximum likelihood panel in Table 4.3 shows the standard error estimates reported by AMOS. (The coefficients are identical those obtained from the two-step method). With one exception (POV), the maximum likelihood standard errors are all somewhat lower than those obtained with the two-step method (with a specified sample size of 431).

MULTIPLE IMPUTATION

Although maximum likelihood is an excellent method for handling missing data, it does have limitations. The principal limitation is that one must specify a joint probability distribution for all the variables, and such models are not always easy to come by. Consequently, although models and software are readily available in the linear and log-linear cases, there is no commercial software for maximum likelihood with missing data for logistic regression, Poisson regression or Cox regression.

An excellent alternative is multiple imputation (Rubin, 1987), which has statistical properties that are nearly as good as maximum likelihood. Like maximum likelihood, multiple imputation estimates are consistent and asymptotically normal. They are close to being asymptotically efficient. (In fact, you can get as close as you like by having a sufficient number of imputations.) Like maximum likelihood, multiple imputation has these desirable properties under either the MAR assumption or a correctly specified model for the missing-data mechanism. However, most software assumes MAR.

Compared with maximum likelihood, multiple imputation has two big advantages. First, it can be applied to virtually any kind of data or model. Second, the analysis can be done using conventional software rather than having to use a special package like LEM or AMOS. The major downside of multiple imputation is that it produces different results every time you use it. That is because the imputed values are random draws rather than deterministic quantities. A second downside is that there are many different ways to do multiple imputation, possibly leading to uncertainty and confusion.

The most widely-used method for multiple imputation is the Markov Chain Monte Carlo (MCMC) algorithm based on

linear regression. This method was first implemented in the stand-alone package NORM (Schafer, 1997), but is now available in SAS and S-PLUS. The approach is quite similar to the multivariate normal EM algorithm which, as we saw earlier, is equivalent to iterated linear regression imputation. There is one major difference, however. After generating predicted values based on the linear regressions, random draws are made from the (simulated) error distribution for each regression equation. These random 'errors' are added to the predicted values for each individual to produce the imputed values. The addition of this random variation compensates for the downward bias in variance estimates that usually results from deterministic imputation methods.

If you apply conventional analysis software to a single data set produced by this random imputation method, you get parameter estimates that are approximately unbiased. However, standard errors will still be underestimated because, as noted earlier, the software can not distinguish real values from imputed values, and the imputed values contain much less information. The parameter estimates will also be inefficient because random variation in the imputed values induces additional sampling variability.

The solution to both of these problems is to do the imputation more than once. Specifically, create several data sets, each with different, randomly drawn, imputed values. If we then apply conventional software to each data set, we get several sets of alternative estimates. These may be combined into a single set of parameter estimates and standard errors using two simple rules (Rubin, 1987). For parameter estimates, one simply takes the mean of the estimates over the several data sets. Combining the standard errors is a little more complicated. First, take the average of the squared standard errors across the several data sets. This is the 'within' variance. The 'between' variance is just the sample variance of the parameter estimates across the several data sets. Add the within and between variances (applying a small correction factor to the latter) and take the square root. The formula is as follows:

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

M is the number of data sets, s_k is the standard error in the kth data set, and b_k is the parameter estimated in the kth data set.

There is one further complication to the method. For this standard error formula to be accurate, the regression parameters used to generate the predicted values must themselves be random draws from their 'posterior' distribution, one random draw for each data set. Otherwise, there will be insufficient variation across data sets. For details, see Schafer (1997).

How many data sets are necessary for multiple imputation? With moderate amounts of missing data, five data sets (the default in SAS) are usually sufficient to get parameter estimates that are close to being fully efficient. Somewhat more may be necessary to get sufficiently stable p-values and confidence intervals. More data sets may also be needed if the fraction of missing data is large.

For the NLSY example, I used PROC MI in SAS to generate 15 'completed' data sets. For each data set, I used PROC REG to estimate the linear regression of ANTI on the other variables. Finally, I used PROC MIANALYZE to combine the results into a single set of parameter estimates, standard errors, confidence intervals and p-values. While this may seem like a lot of work, the programming is really quite simple. Here is the SAS program that accomplished all of these tasks:

```
proc mi data=nlsy out=miout
    nimpute=15;
var anti self pov black hispanic
    divorce gender momwork;
proc reg data=miout outest=a covout;
model anti=self pov black hispanic
    divorce gender momwork;
by _imputation_;
proc mianalyze data=a;
```

```
var intercept self pov black
    hispanic divorce gender momwork;
run;
```

PROC MI reads the NLSY data set and produces a new data set called MIOUT. This data set actually consists of 15 stacked-data sets, with a variable named _IMPUTATION_ having values 1 through 15 to distinguish the different data sets. The VAR statement specifies the variables that go into the imputation process. Each variable with missing data is imputed using a linear regression of that variable on all the other variables.

PROC REG estimates the desired regression model using the MIOUT data set. The BY statement requests that separate regressions be estimated for each value of _IMPUTATION_. The OUTEST option writes the coefficient estimates to a data set called A and the COVOUT option includes the estimated covariance matrix in that data set. This data set is passed to PROC MIANALYZE, which then applies the combining rules to the each of the coefficients specified in the VAR statement. Clearly, there is a major advantage in being able to do the imputation, the analysis and the combination within a single software package. With a stand-alone imputation program, moving the necessary data sets back and forth between packages can get very tedious.

Results are shown in the last two columns of Table 4.3. As expected, both coefficients and standard errors are very similar to those produced by direct maximum likelihood. Keep in mind that this is just one set of possible estimates produced by multiple imputation. The first panel of Table 4.4 contains another set of estimates produced by the same SAS program.

COMPLICATIONS

Space is not sufficient for a thorough treatment of various complications that may arise in the application of multiple imputation. However, it is certainly worth mentioning some of the more important issues that frequently arise.

Auxiliary variables

An auxiliary variable is one that is used in the imputation process but does not appear in the model to be estimated. The most desirable auxiliary variables are those that are moderately to highly correlated with the variables having missing data. Such variables can be very helpful in getting more accurate imputations, thereby increasing the efficiency of the parameter estimates. If auxiliary variables are also associated with the probability that other variables are missing, their inclusion can also reduce bias. In fact, including such variables can go a long way toward making the MAR assumption more plausible.

The dependent variable

If the goal is to estimate some kind of regression model, two questions arise regarding the dependent variable. First, should the

Table 4.4 Regression of ANTI using two multiple imputation methods

	Multivariate normal MCMC		*Sequential generalized regression*	
	Coeff.	*SE*	*Coeff.*	*SE*
SELF	**–0.065**	0.021	**–0.067**	0.021
POV	**0.635**	0.180	**0.700**	0.161
BLACK	0.082	0.160	0.042	0.160
HISPANIC	−0.321	0.173	−0.334	0.173
DIVORCE	−0.112	0.147	−0.129	0.148
GENDER	**–0.553**	0.118	**–0.559**	0.118
MOMWORK	0.235	0.135	0.217	0.157

Coefficients (Coeff.) in bold are statistically significant at the .01 level.
MCMC, Markov Chain Monte Carlo; SE, standard error.

dependent variable be included among the variables used to impute missing values on the independent variables? In conventional, deterministic imputation, the answer is no. Using the dependent variable to impute independent variables can lead to overestimates of the magnitudes of the coefficients. With multiple imputation, however, the answer is definitely yes, because the random component avoids any bias. In fact, leaving out the dependent variable will yield regression coefficients that are attenuated toward zero (Landerman et al., 1997).

Second, should the dependent variable itself be imputed? If the data are MAR and there are no auxiliary variables, the answer is no. Imputation of the dependent variable merely increases sampling variability (Little, 1992). So the preferred procedure is to delete cases with missing data on the dependent variable before doing the imputation. If there are auxiliary variables that are strongly correlated with the dependent variable, imputation of the dependent variable can be helpful in increasing efficiency and, in some cases, reducing bias. Often, one of the best auxiliary variables is the same variable measured at a different point in time.

Combining test statistics

With multiple imputation, any parameter estimates can simply be averaged over the multiple data sets. But test statistics should *never* be averaged. That goes for t-statistics, z-statistics, chi-square statistics and F-statistics. Special procedures are required for combining hypothesis tests from multiple data sets. These procedures can be based on Wald tests, likelihood ratio tests, or a simple method for combining chi-square statistics. For details, see Schafer (1997) or Allison (2001).

Model congeniality

Any multiple imputation method must be based on some model for the data (the imputation model), and that model is not necessarily (or even usually) the same model that one desires to estimate (the analysis model). That raises the question of how similar the imputation model and the analysis model must be in order to get good results. Although they do not have to be identical, the two models should be 'congenial' in the sense that the imputation model should be able to reproduce the major features of the data that are the object of the analysis model (Rubin, 1987; Meng, 1994). Trouble is most likely to occur if the imputation model is simpler than the analysis model. Two examples:

1. The analysis model treats a variable as categorical but the imputation model treats it as quantitative.
2. The analysis model includes interactions and nonlinearities, but the imputation model is strictly linear.

If the fraction of missing data is small, this lack of congeniality may be unproblematic. But if the fraction of missing data is large, results may be misleading.

One implication of the congeniality principle is that imputation models should be relatively 'rich' so that they may be congenial with lots of different models that could be of interest. However, there are serious practical limitations to the complexity of the imputation model. And if the imputer and analyst are different people, it may be quite difficult for the imputer to anticipate the kinds of models that will be estimated with the data. Consequently, it may often be necessary (or at least desirable) to produce different imputed data sets for different analysis models.

Longitudinal data

Longitudinal studies are particularly prone to missing data because subjects often drop out, die, or cannot be located. While there are many kinds of longitudinal data, I focus here on the most common kind, often referred to as panel data. In panel data, one or more variables are measured repeatedly, and the measurements are taken at the same times for all subjects.

Missing data in panel studies can be readily handled by the methods of maximum

likelihood and multiple imputation that we have already discussed. For multiple imputation, the critical consideration is that the imputation must be done in such a way that it reproduces the correlations over time. This is most easily accomplished if the data are formatted so that there is only one record per subject rather than separate records for each observation time point. The imputation model should be formulated so that each variable with missing data may be imputed based on any of the variables at any of the time points (including the variable itself at a different time point).

Categorical variables

The MCMC method based on the multivariate-normal model is the most popular approach to multiple imputation for good reasons. It can handle virtually any pattern of missing data, and it is extremely efficient computationally. Its biggest disadvantage, however, is that it presumes that every variable with missing data is normally distributed and that is clearly not the case for categorical variables. I ignored this problem for the NLSY example, treating each categorical variable as a set of dummy variables and imputing the dummies just like any other variables.

Of course, the resulting imputed values for the dummy variables can be any real numbers and not infrequently, are greater than 1 or less than 0. Many authorities (including me in my 2001 book) recommend rounding the imputed values to 0 and 1 before estimating the analysis model (Schafer, 1997). However, recent analytical and simulation results suggest that this nearly always makes things worse (Horton et al., 2003; Allison 2006). If the dummy variables are to be used as predictor variables in some kind of regression analysis, you are better off just leaving the imputed values as they are. For categorical variables with more than two categories, there is no need to attempt to impose consistency on the imputed values for the multiple dummy variables. Unless the fraction of cases in any one category is very small, this approach usually produces good results.

Alternative methods may be necessary if the fraction of cases in a category is very small (say, 5% or less), or if the analysis method requires that the imputed variable be truly categorical (e.g., the imputed variable is the dependent variable in a logistic regression). In the next section, we will consider some methods more appropriate for the imputation of categorical variables.

OTHER IMPUTATION METHODS

There are numerous alternative models and computational methods that are available for doing multiple imputation. One class of methods uses the MCMC algorithm but applies it to models other than the multivariate normal model. For example, Schafer (http://www.stat.psu.edu/~jls) has developed a freeware package called CAT (available only as an S-Plus library), which is designed for data in which all the variables are categorical. It uses the MCMC algorithm under a multinomial model or a restricted log-linear model.

Schafer has another package called MIX (also available only for S-Plus) that is suitable for data sets and models that include both categorical and quantitative variables. The model is a multinomial (or restricted log-linear) model for the categorical data. Within each cell of the contingency table formed by the categorical variables, the quantitative variables are assumed to follow a multivariate normal distribution with means that may vary across cells but a covariance matrix that is constant across cells. While this method might seem to be ideal for many situations, the model is rather complex and requires considerable thought and care in its implementation.

It is also possible to do imputation under the multivariate normal model but with an algorithm other than MCMC to produce the imputed values. AMELIA, for example, is a stand-alone package that uses the SIR (sampling/importance resampling) algorithm. This is a perfectly respectable approach.

Indeed, the authors claim that it is more computationally efficient than the MCMC algorithm (King et al., 1999).

Perhaps the most promising alternative method for multiple imputation is an approach that is described as either 'sequential generalized regression' or 'multiple imputation for chained equations' (MICE). Instead of assuming a single multivariate model for all the data, one specifies a separate regression model that is used to impute each variable with missing data. Typically, this is a linear regression model for quantitative variables, a logistic regression model (either binomial or multinomial) for categorical variables or a Poisson regression model for count variables (Brand, 1999; Raghunathan et al., 2001).

These models are estimated sequentially using available data, starting with the variable that has the least missing data and proceeding to the variable with the most missing data. After each model is estimated, it is used to generate imputed values for the missing data. For example, in the case of logistic regression, the model is applied to generate predicted probabilities of falling into each category for each case with missing data. These probabilities are then used as the basis for making random draws from the possible values of the categorical variable.

Once imputed values have been generated for all the missing data, the sequential imputation process is repeated, except now the imputed values of the previous round are used as predictors for imputing other variables. This is one thing that distinguishes sequential generalized regression from the MCMC algorithm – in the latter, values imputed for one variable are never used as predictors to impute other variables. The sequential process is repeated for many rounds, with a data set selected at periodic intervals, say, every tenth round.

As noted, the main attraction of sequential generalized regression methods (compared with MCMC methods) is that it is unnecessary to specify a comprehensive model for the joint distribution of all the variables. Potentially, then, one can tailor the imputation model to be optimally suited for each variable that has missing data. A major disadvantage is that, unlike MCMC, there is no theory that guarantees that the sequential method will converge to the correct distribution for the missing values. Recent simulation studies suggest that the method works well, but such studies have only examined a limited range of circumstances (Van Buuren et al., 2006). The sequential method may also require a good deal more computing time, simply because estimation of logistic and Poisson models is more intensive than estimation of linear models.

User contributed add-ons for sequential generalized regression are currently available for SAS (Raghunathan et al., 2000), S-Plus (Van Buuren and Oudshoorn, 2000), and Stata (Royston, 2004). In the remainder of this section, I apply the ICE command for Stata to the NLSY data set. Here are the Stata commands:

```
use "d:\nlsy.dta"
gen race=1+black+2*hispanic
ice anti self pov race divorce
    gender momwork, dryrun
ice anti self pov race black
    hispanic divorce gender momwork
    using nlsyimp,
    m(15) passive(black:race==2\
    hispanic:race==3)
    substitute(race:black hispanic)
use nlsyimp, clear
micombine regress anti self pov
    black hispanic divorce gender
    momwork
```

A bit of explanation is needed here. The GEN command creates a new variable RACE that has values of 1, 2 or 3, corresponding to white/non-Hispanic, black and Hispanic. It is better to impute this variable rather than the individual dummies for black and Hispanic because that ensures that each person with missing race data will be assigned to one and only one category.

The first ICE command is a 'dryrun'. It scans the data set, identifies the variables with missing data, and proposes an imputation model for each one. In this case, ICE proposed a linear model for imputing

SELF, binary-logit models for imputing POV and MOMWORK, and a multinomial-logit model for imputing RACE. The second ICE command actually does the imputation, using the default methods that were proposed, and writes the imputed data sets into the single Stata data set, NLSYIMP. The M(15) option requests 15 data sets, distinguished by the variable _j, which has values of 1 through 15. The PASSIVE option says that the dummy variables BLACK and HISPANIC are imputed 'passively' based on the imputed values of RACE. The SUBSTITUTE option tells ICE to use the dummy variables BLACK and HISPANIC as predictors when imputing other variables, rather than the 3-category variable RACE. Without this option, RACE would be treated as a quantitative predictor which would clearly be inappropriate.

The USE command switches from the original data set to the newly-imputed data set. The MICOMBINE command (along with the REGRESS command) estimates the regression model for the 15 imputed data sets and then combines the results into a single set of estimates and test statistics. Results are shown in the second panel Table 4.4. The first panel of this table is simply a replication of the MCMC multivariate normal method that produced the results in the last panel of Table 4.3. It is included here for comparison with the sequential generalized regression results, but also to illustrate the degree to which results may vary from one replication of multiple imputation to another. Although the coefficients vary slightly across the replications and methods, they all tell essentially the same story. And the differences between the MCMC results and the sequential results are no greater than the differences between one run of MCMC and another.

SUMMARY AND CONCLUSION

Conventional methods for handling missing data are seriously flawed. Even under the best of conditions, they typically yield biased parameter estimates, biased standard error estimates, or both. Despite the often substantial loss of power, listwise deletion is probably the safest method because it is not prone to Type I errors. On the other hand, conventional imputation methods may be the most dangerous because they often lead to serious underestimates of standard errors and *p*-values.

By contrast, maximum likelihood and multiple imputation have nearly optimal statistical properties and they possess these properties under assumptions that are typically weaker than those used to justify conventional methods. Specifically, maximum likelihood and multiple imputation perform well under the assumption that the data are MAR, rather than the more severe requirement of MCAR; and if the data are not MAR, these two methods do well under a correctly specified model for missingness (something that is not so easy to come by).

Of the two methods, I prefer maximum likelihood because it yields a unique set of estimates, while multiple imputation produces different results every time you use it. Software for maximum likelihood estimation of linear models with missing data is readily available in most stand-alone packages for linear-structural equation modeling, including LISREL, AMOS, EQS and M-PLUS. For log-linear modeling of categorical data, there is the freeware package LEM.

Multiple imputation is an attractive alternative when estimating models for which maximum likelihood is not currently available, including logistic regression and Cox regression. It also has the advantage of not requiring the user to master an unfamiliar software package to do the analysis. The downside, of course, is that it does not produce a determinate result. And there are lots of different ways to do multiple imputation, so some care must go into choosing the most suitable method for a particular application.

Both maximum likelihood and multiple imputation usually require more time and effort than conventional methods for handling missing data. With improvements in software, however, both methods have become much easier to implement, and further

improvements can be expected. And sometimes you just have to do more to get things right. Nowadays, there is no good excuse for avoiding these clearly superior methods.

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

Arbuckle, J.L. (1996) 'Full information estimation in the presence of incomplete data', in Marcoulides, G.A. and Schumacker, R.E. (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, Erasmus University Rotterdam.

Cohen, J. and Cohen, P. (1985) *Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences* (2nd edn.). Mahwah, NJ: Lawrence Erlbaum Associates.

Dempster, A.P., Laird, N.M. and Donald R.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., S.M. Hofer and 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. and 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.

King, G., Honaker, J., Joseph, A., Scheve, K. and Singh, N. (1999) 'AMELIA: A program for missing data.' Unpublished program manual. Online. Available: http://gking.harvard.edu/stats.shtml.

Landerman, L.R., Land, K.C. and Pieper, C.F. (1997) 'An empirical evaluation of the predictive mean matching method for imputing missing values', *Sociological Methods and Research* 26: 3–33.

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. and Rubin, D.B. (2002). *Statistical analysis with missing data* (2nd edn.). New York: Wiley.

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

Raghunathan T.E., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P. (2001) 'A multivariate technique for multiply imputing missing valuesusing a sequence of regression models', *Survey Methodology*, 27: 85–95.

Raghunathan, T.E., Solenberger, P. and van Hoewyk, J. (2000) 'IVEware: imputation and variance estimation software: installation instructions and user guide.' Survey Research Center, Institute of Social Research, University of Michigan. Online. Available: http://www.isr.umich.edu/src/smp/ive/

Robins, J.M. and 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. and 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 and Hall.

Scharfstein, D.O., Rotnizky, A. and Robins, J.M. (1999) 'Adjusting for nonignorable drop-out using semiparametric nonresponse models (with comments)', *Journal of the American Statistical Association* 94: 1096–1146.

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

Van Buuren, S. and Oudshoorn, C.G.M. (2000) 'Multivariate imputation by chained equations: MICE V1.0 user's manual.' *Report* PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid.

Van Praag, B.M.S., Dijkstra, T.K. and Van Velzen, J. (1985) 'Least-squares theory based on general distributional assumptions with an application to the incomplete observations problem', *Psychometrika* 50: 25–36.