For information:

Sage Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

Sage Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

Sage Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Allison, Paul David.
Missing data / by Paul D. Allison.
p. cm. – (Quantitative applications in the social sciences (Qass) ; no. 07-136)
Includes bibliographical references and index.
ISBN 0-7619-1672-5 (p)
1. Mathematical statistics. 2. Missing observations (Statistics)
I. Title. II. Series: Sage university papers series. Quantitative applications in the social sciences; no. 07-136
QA276 .A55 2001
001.4'22–dc21 2001001295

This book is printed on acid-free paper.

02 03 04 05 06 07 10 9 8 7 6 5 4 3 2 1

Acquiring Editor: C. Deborah Laughton
Editorial Assistant: Eileen Carr
Production Editor: Denise Santoyo
Production Assistant: Kathryn Journey
Typesetter: Technical Typesetting Inc.

When citing a university paper, please use the proper form. Remember to cite the Sage University Paper series title and include paper number. One of the following formats can be adapted (depending on the style manual used):

(1) ALLISON, P. D. (2001) *Missing Data.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

OR

(2) Allison, P. D. (2001). *Missing Data.* (Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-136). Thousand Oaks, CA: Sage.

CONTENTS

SERIES EDITOR'S INTRODUCTION

Problems of missing data are pervasive in empirical social science research. The statistical results reported with most nonexperimental studies rest on sample sizes smaller, sometimes much smaller, than the initial number of selected cases. A relatively few absent observations on a handful of variables can quickly reduce the effective N. With an opinion survey for example, it is not uncommon for a multivariate analysis to halve the original draw. Suppose Professor Mary Rose, of the Business School, is examining a probability sample of $N = 1{,}000$ respondents, in a survey of consumer attitudes and behavior. She estimates a reasonably specified multiple regression model of spending, employing the usual computer option of *listwise deletion* (i.e., any respondent with data lacking on any model variable is excluded). As a result, the actual cases available fall to $N = 499$. Serious questions arise. Do these 499 still "represent" the population? Are the coefficients possessing of any desirable properties? Is the sample too small for rejection of null hypotheses? In order to keep sample size, should pairwise deletion have been tried? Or, are there altogether new approaches worth considering? These questions, and others, are addressed in this splendid monograph by Paul Allison.

"Observations are randomly missing." That is the stock argument for going ahead in the face of data attrition, relying on the cases left. But the assumption is vague, and may not be saving. Suppose the observations are "missing completely at random" (labeled MCAR by Allison)? That means that none of the variables, dependent (Y) or independent (X), has missing scores related to the values of the variable itself. For example, with the spending variable above, nonresponse should be no more likely for big spenders than small spenders. Given the same condition held for the other model variables, then the subsample of 499 would represent a scientific draw, permitting valid inferences. In particular, it allows the regression estimates to be unbiased and consistent. This variety of randomness, MCAR, is the problem-free sort researchers may like to claim, but it makes very strong assumptions.

disadvantage that is apparent to anyone who has used it: In many applications, listwise deletion can exclude a large fraction of the original sample. For example, suppose you have collected data on a sample of 1,000 people and you want to estimate a multiple regression model with 20 variables. Each of the variables has missing data on 5% of the cases, and the chance that data are missing for one variable is independent of the chance that it is missing on any other variable. You could then expect to have complete data for only about 360 of the cases, discarding the other 640. If you merely downloaded the data from a web site, you might not feel too bad about this, although you might wish you had a few more cases. On the other hand, if you had spent $200 per interview for each of the 1,000 people, you might have serious regrets about the $130,000 that was wasted (at least for this analysis). Surely there must be some way to salvage something from the 640 incomplete cases, many of which may lack data on only one of the 20 variables.

Many alternative methods have been proposed, and several of them will be reviewed in this book. Unfortunately, most of these methods have little value, and many of them are inferior to listwise deletion. That's the bad news. The good news is that statisticians have developed two novel approaches to handling missing data—maximum likelihood and multiple imputation—that offer substantial improvements over listwise deletion. Although the theory behind these methods has been known for at least a decade, it is only in the last few years that they have become computationally practical. Even now, multiple imputation or maximum likelihood can demand a substantial investment of time and energy, both in learning the methods and in carrying them out on a routine basis. But, if you want to do things right, you usually have to pay a price.

Both maximum likelihood and multiple imputation have statistical properties that are about as good as we can reasonably hope to achieve. Nevertheless, it is essential to keep in mind that these methods, like all the others, depend on certain easily violated assumptions for their validity. Not only that, there is no way to test whether or not the most crucial assumptions are satisfied. The upshot is that although some missing data methods are clearly better than others, none of them really can be described as good. The only really good solution to the missing data problem is not to have any. So in the design and execution of research projects, it is essential to put great effort into

minimizing the occurrence of missing data. Statistical adjustments can never make up for sloppy research.

2. ASSUMPTIONS

Researchers often try to make the case that people who have missing values on a particular variable are no different from those with observed measurements. It is common, for example, to present evidence that people who do and do not report their income are not significantly different on a variety of other variables. More generally, researchers have often claimed or assumed that their data are "missing at random" without a clear understanding of what that means. Even statisticians were once vague or equivocal about this notion. However, Rubin (1976) put things on a solid foundation by rigorously defining different assumptions that might plausibly be made about missing data mechanisms. Although his definitions are rather technical, I will try to convey an informal understanding of what they mean.

Missing Completely at Random

Suppose there are missing data on a particular variable Y. The data on Y are said to be *missing completely at random* (MCAR) if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variables in the data set. When this assumption is satisfied for all variables, the set of individuals with complete data can be regarded as a simple random subsample from the original set of observations. Note that MCAR does allow for the possibility that "missingness" on Y is related to "missingness" on some other variable X. For example, even if people who refuse to report their age invariably refuse to report their income, it is still possible that the data could be missing completely at random.

The MCAR assumption *would be* violated if people who did not report their income were younger, on average, than people who did report their income. It would be easy to test this implication by dividing the sample into those who did and did not report their income, and then testing for a difference in mean age. If there are, in fact, no systematic differences on the fully observed variables between those with data present and those with missing data, then the data are said

to be *observed at random*. On the other hand, just because the data pass this test does not mean that the MCAR assumption is satisfied. Still there must be no relationship between missingness on a particular variable and the values of that variable.

Although MCAR is a rather strong assumption, there are times when it is reasonable, especially when data are missing as part of the research design. Such designs are often attractive when a particular variable is very expensive to measure. The strategy then is to measure the expensive variable only for a random subset of the larger sample, implying that data are missing completely at random for the remainder of the sample.

Missing at Random

A considerably weaker assumption is that the data are *missing at random* (MAR). Data on Y are said to be missing at random if the probability of missing data on Y is unrelated to the value of Y, after controlling for other variables in the analysis. To express this more formally, suppose there are only two variables X and Y, where X always is observed and Y sometimes is missing. MAR means that

$$\Pr(Y\text{missing}|Y, X) = \Pr(Y\text{missing}|X).$$

In words, this expression means that the conditional probability of missing data on Y, given both Y and X, is equal to the probability of missing data on Y given X alone. For example, the MAR assumption would be satisfied if the probability of missing data on income depended on a person's marital status, but within each marital status category, the probability of missing income was unrelated to income. In general, data are *not* missing at random if those individuals with missing data on a particular variable tend to have lower (or higher) values on that variable than those with data present, controlling for other observed variables.

It is impossible to test whether the MAR condition is satisfied, and the reason should be intuitively clear. Because we do not know the values of the missing data, we can not compare the values of those with and without missing data to see if they differ systematically on that variable.

Ignorable

The missing data mechanism is said to be ignorable if (a) the data are MAR and (b) the parameters that govern the missing data process are unrelated to the parameters to be estimated. Ignorability basically means that there is no need to model the missing data mechanism as part of the estimation process. However, special techniques certainly are needed to utilize the data in an efficient manner. Because it is hard to imagine real-world applications where condition (b) is not satisfied, I treat MAR and ignorability as equivalent conditions in this book. Even in the rare situation where condition (b) is not satisfied, methods that assume ignorability work just fine, but you could do even better by modeling the missing data mechanism.

Nonignorable

If the data are not MAR, we say that the missing data mechanism is nonignorable. In that case, usually the missing data mechanism must be modeled to get good estimates of the parameters of interest. One widely used method for nonignorable missing data is Heckman's (1976) two-stage estimator for regression models with selection bias on the dependent variable. Unfortunately, for effective estimation with nonignorable missing data, *very* good prior knowledge about the nature of the missing data process usually is needed, because the data contain no information about what models would be appropriate and the results typically will be very sensitive to the choice of model. For these reasons and because models for nonignorable missing data typically must be quite specialized for each application, this book puts the major emphasis on methods for ignorable missing data. In the last chapter, I briefly survey some approaches to handling nonignorable missing data. In Chapter 3, we will see that listwise deletion has some very attractive properties with respect to certain kinds of nonignorable missing data will be evident.

3. CONVENTIONAL METHODS

Although many different methods have been proposed for handling missing data, only a few have gained widespread popularity. Unfortunately, none of the widely used methods is clearly superior to

listwise deletion. In this section, I briefly review some of these methods, starting with the simplest. In evaluating these methods, I will be particularly concerned with their performance in regression analysis (including logistic regression and Cox regression), but many of the comments also apply to other types of analysis as well.

Listwise Deletion

As already noted, listwise deletion is accomplished by deleting from the sample any observations that have missing data on any variables in the model of interest and then applying conventional methods of analysis for complete data sets. There are two obvious advantages to listwise deletion: (1) it can be used for any kind of statistical analysis, from structural equation modeling to log-linear analysis; (2) no special computational methods are required. Depending on the missing data mechanism, listwise deletion also can have some attractive statistical properties. Specifically, if the data are MCAR, then the reduced sample will be a random subsample of the original sample. This implies that, for any parameter of interest, if the estimates would be unbiased for the full data set (with no missing data), they also will be unbiased for the listwise deleted data set. Furthermore, the standard errors and test statistics obtained with the listwise deleted data set will be just as appropriate as they would have been in the full data set.

Of course, the standard errors generally will be larger in the listwise deleted data set because less information is utilized. They also will tend to be larger than standard errors obtained from the optimal methods described later in this book, but at least you do not have to worry about making inferential errors because of the missing data—a big problem with most of the other commonly used methods.

On the other hand, if the data are not MCAR, but only MAR, listwise deletion can yield biased estimates. For example, if the probability of missing data on schooling depends on occupational status, regression of occupational status on schooling will produce a biased estimate of the regression coefficient. So, in general, it appears that listwise deletion is not robust to violations of the MCAR assumption. Surprisingly, however, listwise deletion is the method that is *most* robust to violations of MAR among *independent* variables in a regression analysis. Specifically, if the probability of missing data on any of the independent variables does *not* depend on the values of the *dependent* variable, then regression estimates using listwise deletion

will be unbiased (if all the usual assumptions of the regression model are satisfied).[1]

For example, suppose that we want to estimate a regression model to predict annual savings. One of the independent variables is income, for which 40% of the data are missing. Suppose further that the probability of missing data on income is highly dependent on both income and years of schooling, another independent variable in the model. As long as the probability of missing income does not depend on *savings*, the regression estimates will be unbiased (Little, 1992).

Why is this the case? Here is the essential idea. It is well-known that disproportionate stratified sampling on the independent variables in a regression model does not bias coefficient estimates. A missing data mechanism that depends only on the values of the independent variables is essentially equivalent to stratified sampling, that is, cases are being selected into the sample with a probability that depends on the values of those variables. This conclusion applies not only to linear regression models, but also to logistic regression, Cox regression, Poisson regression, and so on.

In fact, for logistic regression, listwise deletion gives valid inferences under even broader conditions. If the probability of missing data on any variable depends on the value of the dependent variable but does *not* depend on any of the independent variables, then logistic regression with listwise deletion yields consistent estimates of the slope coefficients and their standard errors (Vach, 1994). The intercept estimate will be biased, however. Logistic regression with listwise deletion is problematic only when the probability of any missing data depends on *both* the dependent and independent variables.[2]

To sum up, listwise deletion is not a *bad* method for handling missing data. Although it does not use all of the available information, at least it gives valid inferences when the data are MCAR. As we will see, that is more than can be said for nearly all the other commonplace methods for handling missing data. The methods of maximum likelihood and multiple imputation, discussed in later chapters, are potentially much better than listwise deletion in many situations, but for regression analysis, listwise deletion is even more robust than these sophisticated methods to violations of the MAR assumption. Specifically, whenever the probability of missing data on a particular independent variable depends on the value of that variable (and not the dependent variable), listwise deletion may do better than maximum likelihood or multiple imputation.

There is one important caveat to these claims about listwise deletion for regression analysis. The regression coefficients are assumed to be the same for all cases in the sample. If the regression coefficients vary across subsets of the population, then any nonrandom restriction of the sample (e.g., through listwise deletion) may weight the regression coefficients toward one subset or another. Of course, if such variation in the regression parameters is suspected, either separate regressions should be done in different subsamples or appropriate interactions should be included in the regression model (Winship & Radbill, 1994).

Pairwise Deletion

Also known as available case analysis, pairwise deletion is a simple alternative that can be used for many linear models, including linear regression, factor analysis, and more complex structural equation models. It is well known, for example, that a linear regression can be estimated using only the sample means and covariance matrix or, equivalently, the means, standard deviations, and correlation matrix. The idea of pairwise deletion is to compute each of these summary statistics using all the cases that are available. For example, to compute the covariance between two variables X and Z, all the cases that have data present for both X and Z are used. Once the summary measures have been computed, they can be used to calculate the parameters of interest, for example, regression coefficients.

There are ambiguities in how to implement this principle. When computing a covariance that requires the mean for each variable, do you compute the means using only cases with data on both variables or do you compute them from all the available cases on each variable? There is no point to dwelling on such questions because all the variations lead to estimators with similar properties. The general conclusion is that if the data are MCAR, pairwise deletion produces parameter estimates that are consistent (and, therefore, approximately unbiased in large samples). On the other hand, if the data are only MAR, but not observed at random, the estimates may be seriously biased.

If the data are indeed MCAR, pairwise deletion might be expected to be more efficient than listwise deletion because more information is utilized. By more efficient, I mean that the pairwise estimates would have less sampling variability (smaller true standard errors) than the listwise estimates. This is not always true, however. Both analytical and simulation studies of linear regression models indicate that pairwise deletion produces more efficient estimates when the correlations among the variables are generally low, whereas listwise deletion does better when the correlations are high (Glasser, 1964; Haitovsky, 1968; Kim & Curry, 1977).

The big problem with pairwise deletion is that the estimated standard errors and test statistics produced by conventional software are biased. Symptomatic of that problem is that when you input a covariance matrix to a regression program, you must also specify the sample size to calculate standard errors. Some programs for pairwise deletion use the number of cases on the variable with the most missing data, whereas others use the minimum of the number of cases used to compute each covariance. No single number is satisfactory, however. In principle, it is possible to get consistent estimates of the standard errors, but the formulas are complex and have not been implemented in any commercial software.[3]

A second problem that occasionally arises with pairwise deletion, especially in small samples, is that the constructed covariance or correlation matrix may not be "positive definite," which implies that the regression computations cannot be carried out at all. Because of these difficulties, as well as its relative sensitivity to departures from MCAR, pairwise deletion cannot be generally recommended as an alternative to listwise deletion.

Dummy Variable Adjustment

There is another method for missing predictors in a regression analysis that is remarkably simple and intuitively appealing (Cohen & Cohen, 1985). Suppose that some data are missing on a variable X, which is one of several independent variables in a regression analysis. We create a dummy variable D that is equal to 1 if data are missing on X and equal to 0 otherwise. We also create a variable X^* such that

$$X^* = \begin{cases} X & \text{when data are not missing,} \\ c & \text{when data are missing,} \end{cases}$$

where c can be any constant. We then regress the dependent variable Y on X^*, D, and any other variables in the intended model. This

technique, known as dummy variable adjustment or the missing-indicator method, can be extended easily to the case of more than one independent variable with missing data.

The apparent virtue of the dummy variable adjustment method is that it uses all the information that is available about the missing data. Substitution of the value c for the missing data is not properly regarded as imputation because the coefficient of X^* is invariant to the choice of c. Indeed, the only aspect of the model that depends on the choice of c is the coefficient of D, the missing value indicator. For ease of interpretation, a convenient choice of c is the mean of X for nonmissing cases. Then the coefficient of D can be interpreted as the predicted value of Y for individuals with missing data on X minus the predicted value of Y for individuals at the mean of X, controlling for other variables in the model. The coefficient for X^* can be regarded as an estimate of the effect of X among the subgroup of those that have data on X.

Unfortunately, this method generally produces biased estimates of the coefficients, as proven by Jones (1996).[4] A simple simulation illustrates the problem. I generated 10,000 cases on three variables, X, Y, and Z, by sampling from a trivariate normal distribution. For the regression of Y on X and Z, the true coefficients for each variable were 1.0. For the full sample of 10,000, the least-squares regression coefficients, shown in the first column of Table 3.1 are—not surprisingly—quite close to the true values.

I then randomly made some of the Z values missing with a probability of 1/2. Because the probability of missing data is unrelated to any other variable, the data are MCAR. The second column in Table 3.1 shows that listwise deletion yields estimates that are very close to those obtained when no data are missing. On the other hand, the coefficients for the dummy variable adjustment method are

TABLE 3.1
Regression in Simulated Data for Three Methods

Coefficient of	*Full Data*	*Listwise Deletion*	*Dummy Variable Adjustment*
X	0.98	0.96	1.28
Z	1.01	1.03	0.87
D			0.02

clearly biased—too high for the X coefficient and too low for the Z coefficient.

A closely related method has been proposed for categorical independent variables in regression analysis. Such variables are typically handled by creating a set of dummy variables, one variable for each of the categories except for a reference category. The proposal is simply to create an additional category—and an additional dummy variable—for those individuals with missing data on the categorical variables. Again, however, we have an intuitively appealing method that is biased even when the data are MCAR (Jones, 1996; Vach and Blettner, 1991).

Imputation

Many missing data methods fall under the general heading of imputation. The basic idea is to substitute some reasonable guess (imputation) for each missing value and then proceed to do the analysis as if there were no missing data. Of course, there are lots of different ways to impute missing values. Perhaps the simplest is marginal mean imputation: For each missing value on a given variable, substitute the mean for those cases with data present on that variable. This method is well known to produce biased estimates of variances and covariances (Haitovsky, 1968) and generally should be avoided.

A better approach is to use information on other variables by way of multiple regression, a method sometimes known as conditional mean imputation. Suppose we are estimating a multiple regression model with several independent variables. One of those variables, X, has missing data for some of the cases. For those cases with complete data, we regress X on all the other independent variables. Using the estimated equation, we generate predicted values for the cases with missing data on X. These are substituted for the missing data and the analysis proceeds as if there were no missing data.

The method gets more complicated when more than one independent variable has missing data, and there are several variations on the general theme. In general, if imputations are based solely on other independent variables (not the dependent variable) and if the data are MCAR, the least-squares coefficients are consistent, implying that they are approximately unbiased in large samples (Gourieroux & Monfort, 1981). However, they are not fully efficient. Improved esti-

mators can be obtained using weighted least squares (Beale & Little, 1975) or generalized least squares (Gourieroux & Monfort, 1981).

Unfortunately, all of these imputation methods suffer from a fundamental problem: Analyzing imputed data as though it were complete data produces standard errors that are underestimated and test statistics that are overestimated. Conventional analytic methods simply do not adjust for the fact that the imputation process involves uncertainty about the missing values.[5] In later chapters, an approach to imputation that overcomes these difficulties is examined.

Summary

All the common methods for salvaging information from cases with missing data typically make things worse: They introduce substantial bias, make the analysis more sensitive to departures from MCAR, or yield standard error estimates that are incorrect (usually too low). In light of these shortcomings, listwise deletion does not look so bad. However, better methods are available. In the next chapter, maximum likelihood methods that are available for many common modeling objectives are examined. In Chapters 5 and 6, multiple imputation, which can be used in almost any setting, is considered. Both methods have very good properties if the data are MAR. In principle, these methods also can be used for nonignorable missing data, but that requires a correct model of the process by which data are missing—something that usually is difficult to come by.

4. MAXIMUM LIKELIHOOD

Maximum likelihood (ML) is a very general approach to statistical estimation that is widely used to handle many otherwise difficult estimation problems. Most readers will be familiar with ML as the preferred method for estimating the logistic regression model. Ordinary least-squares linear regression is also an ML method when the error term is assumed to be normally distributed. It turns out that ML is particularly adept at handling missing data problems. In this chapter, I begin by reviewing some general properties of ML estimates. Then I present the basic principles of ML estimation under the assumption that the missing data mechanism is ignorable. These principles

are illustrated with a simple contingency table example. The remainder of the chapter considers more complex examples where the goal is to estimate a linear model, based on the multivariate normal distribution.

Review of Maximum Likelihood

The basic principle of ML estimation is to choose as estimates those values that, if true, would maximize the probability of observing what has, in fact, been observed. To accomplish this, we first need a formula that expresses the probability of the data as a function of both the data and the unknown parameters. When observations are independent (the usual assumption), the overall likelihood (probability) for the sample is just the product of all the likelihoods for the individual observations.

Suppose we are trying to estimate a parameter θ. If $f(y|\theta)$ is the probability (or probability density) of observing a single value of Y given some value of θ, the likelihood for a sample of n observations is

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta),$$

where $\prod$ is the symbol for repeated multiplication. Of course, we still need to specify exactly what $f(y|\theta)$ is. For example, suppose Y is a dichotomous variable coded 1 or 0, and θ is the probability that $Y = 1$. Then

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}.$$

Once we have $L(\theta)$—which is called the likelihood function—there are a variety of techniques to find the value of θ that makes the likelihood as large as possible.

ML estimators have a number of desirable properties. Under a fairly wide range of conditions, they are known to be consistent, asymptotically efficient, and asymptotically normal (Agresti & Finlay, 1997). Consistency implies that the estimates are approximately unbiased in large samples. Efficiency implies that the true standard errors are at least as small as the standard errors for any other consistent

estimators. The asymptotic part means that this statement is only approximately true, and the approximation gets better as the sample size gets larger. Finally, asymptotic normality means that in repeated sampling, the estimates have an approximately normal distribution (again, the approximation improves with increasing sample size). This justifies the use of a normal table to construct confidence intervals or compute p values.

ML With Missing Data

What happens when data are missing for some of the observations? When the missing data mechanism is ignorable (and hence MAR), we can obtain the likelihood simply by summing the usual likelihood over all possible values of the missing data. Suppose, for example, that we attempt to collect data on two variables, X and Y, for a sample of n independent observations. For the first m observations, we observe both X and Y, but for the remaining $n-m$ observations, we are only able to measure Y. For a single observation with complete data, let us represent the likelihood by $f(x, y|\theta)$, where θ is a set of unknown parameters that govern the distribution of X and Y. Assuming that X is discrete, the likelihood for a case with missing data on X is just the "marginal" distribution of Y:

$$g(y|\theta) = \sum_x f(x, y|\theta).$$

(When X is continuous, the summation is replaced by an integral.) The likelihood for the entire sample is just

$$L(\theta) = \prod_{i=1}^{m} f(x_i, y_i|\theta) \prod_{i=m+1}^{n} g(y_i|\theta).$$

The problem then becomes one of finding values of θ to make this likelihood as large as possible. A variety of methods are available to solve this optimization problem, and a few of them will be considered later.

ML is particularly easy when the pattern of missing data is *monotonic*. In a monotonic pattern, the variables can be arranged in an order such that for any observation in the sample, if data are missing on a particular variable, they also must be missing for all variables that come later in the order. Here is an example with four variables, X_1, X_2, X_3, and X_4. There are no missing data on X_1. Ten percent of the cases are missing on X_2. Those cases that are missing on X_2 also have missing data on X_3 and X_4. An additional 20% of the cases have missing data on both X_3 and X_4, but not on X_2. A monotonic pattern often arises in panel studies, where people drop out at various points in time and never return.

If only one variable has missing data, the pattern is necessarily monotonic. Consider the two-variable case with data missing on X only. The joint distribution $f(x, y)$ can be written as $h(x|y)g(y)$, where $g(y)$ is the marginal distribution of Y (previously defined) and $h(x|y)$ is the conditional distribution of X given Y. This enables us to rewrite the likelihood as

$$L(\lambda, \phi) = \prod_{i=1}^{m} h(x_i|y_i; \lambda) \prod_{i=1}^{n} g(y_i|\phi).$$

This expression differs from the earlier one in two important ways. First, the second product is over *all* the observations, not just those with missing data on X. Second, the parameters have been separated into two parts: λ describes the conditional distribution of X given Y and ϕ describes the marginal distribution of Y. These changes imply that we can maximize the two parts of the likelihood separately, typically using conventional estimation procedures for each part. Thus, if X and Y have a bivariate normal distribution, we can calculate the mean and variance of Y for the entire sample. Then, for those cases with data on X, we can calculate the regression of X on Y. The resulting parameter estimates can be combined to produce ML estimates for any other parameters we might be interested in, for example, the correlation coefficient.

Contingency Table Data

These features of ML estimation can be illustrated very concretely with contingency table data. Suppose for a simple random sample of 200 people, we attempt to measure two dichotomous variables, X and Y, with possible values of 1 and 2. For 150 cases, we observe both X and Y, and obtain the results shown in the following contingency

```
$Sample size = 1302
$missing = -9
$input variables
  gradrat
  csat
  lenroll
  private
  stufac
  rmbrd
  act
$rawdata
$include=c:\college.dat
$mstructure
  csat
  lenroll
  private
  stufac
  rmbrd
  act
$structure
  gradrat = () + csat + lenroll + private + stufac
            + rmbrd + (1) error
  act<>error
```

Figure 4.1. Amos Commands for the Regression Model That Predicts GRADRAT

estimates are identical, but the Amos standard errors are noticeably larger—which is just what we would expect. They are still quite a bit smaller than those in Table 4.2 that were obtained with listwise deletion.

Conclusion

Maximum likelihood can be an effective and practical method for handling data that are missing at random. In this situation, ML estimates are known to be optimal in large samples. For linear models that fall within the general class of structural equation models estimated by programs like LISREL, ML estimates are easily obtained

TABLE 4.6
Regressions That Predict GRADRAT Using Direct ML With Amos

Variable	*Coefficient*	*Standard Error*	*t Statistic*	*p Value*
INTERCEPT	−32.395	4.863	−6.661	0.000000
CSAT	0.067	0.005	13.949	0.000000
LENROLL	2.083	0.595	3.499	0.000467
PRIVATE	12.914	1.277	10.114	0.000000
STUFAC	−0.181	0.092	−1.968	0.049068
RMBRD	2.404	0.548	4.386	0.000012

by several widely available software packages. Software is also available for ML estimation of log-linear models for categorical data, but the implementation in this setting is somewhat less straightforward. One limitation of the ML approach is that it requires a model for the joint distribution of all variables with missing data. The multivariate normal model is often convenient for this purpose, but may be unrealistic for many applications.

5. MULTIPLE IMPUTATION: BASICS

Although ML represents a major advance over conventional approaches to missing data, it has its limitations. As we have seen, ML theory and software are readily available for linear models and log-linear models, but beyond that, either theory or software is generally lacking. For example, if you want to estimate a Cox proportional hazards model or an ordered logistic regression model, you will have a tough time implementing ML methods for missing data. Even if your model *can* be estimated with ML, you will need to use specialized software that may lack diagnostics or graphical output that you particularly want.

Fortunately, there is an alternative approach—multiple imputation—that has the same optimal properties as ML, but removes some of these limitations. More specifically, multiple imputation (MI), when used correctly, produces estimates that are consistent, asymptotically efficient, and asymptotically normal when the data are MAR. Unlike ML, multiple imputation can be used with virtually any kind of data and any kind of model, and the analysis can be done with unmodified, conventional software. Of course MI has its own drawbacks. It

can be cumbersome to implement and it is easy to do it the wrong way. Both of these problems can be substantially alleviated by using good software to do the imputations. A more fundamental drawback is that MI produces different estimates (hopefully, only slightly different) every time you use it. That can lead to awkward situations in which different researchers get different numbers from the same data using the same methods.

Single Random Imputation

The reason that MI does not produce a unique set of numbers is that random variation is deliberately introduced in the imputation process. Without a random component, deterministic imputation methods generally produce underestimates of variances for variables with missing data and, sometimes, covariances as well. As we saw in the previous chapter, the EM algorithm for the multivariate normal model solves that problem by using residual variance and covariance estimates to correct the conventional formulas. However, a good alternative is to make random draws from the residual distribution of each imputed variable and add those random numbers to the imputed values. Then, conventional formulas can be used to calculate variances and covariances.

Here is a simple example. Suppose we want to estimate the correlation between X and Y, but data are missing on X for, say, 50% of the cases. We can impute values for the missing Xs by regressing X on Y for the cases with complete data and then using the resulting regression equation to generate predicted values for the cases that are missing on X. I did this for a simulated sample of 10,000 cases, where X and Y were drawn from a standard, bivariate normal distribution with a correlation of 0.30. Half of the X values were assigned to be missing (completely at random). Using the predicted values from the regression of X on Y to substitute for the missing values, the correlation between X and Y was estimated to be 0.42.

Why the overestimate? The sample correlation is just the sample covariance of X and Y divided by the product of their sample standard deviations. The regression imputation method yields unbiased estimates of the covariance. Moreover, the standard deviation of Y(with no missing data) was correctly estimated at about 1.0. However, the standard deviation of X (including the imputed values) was only 0.74, whereas the true standard deviation was 1.0, resulting in an overestimate of the correlation. An alternative way to think about the problem is that the imputed value of X for the 5,000 cases with missing data is a perfect linear function of Y, thereby inflating the correlation between the two variables.

We can correct this bias by taking random draws from the residual distribution of X and then adding these random numbers to the predicted values of X. In this example, the residual distribution of X (regressed on Y) is normal with a mean of 0 and a standard deviation (estimated from the listwise deleted least-squares regression) of 0.9525. For case i, let u_i be a random draw from a standard normal distribution and let $\hat{x}_i$ be the predicted value from the regression of X on Y. Our modified imputed value is then $\tilde{x}_i = \hat{x}_i + 0.9525u_i$. For all observations in which X is missing, we substitute $\tilde{x}_i$ and then compute the correlation. When I did this for the simulated sample of 10,000 cases, the correlation between X (with modified imputed values) and Y was 0.316, only a little higher than the true value of 0.300.

Multiple Random Imputation

Random imputation can eliminate the biases that are endemic to deterministic imputation, but a serious problem remains. If we use imputed data (either random or deterministic) as if it were real data, the resulting standard error estimates generally will be too low and test statistics will be too high. Conventional methods for standard error estimation cannot account adequately for the fact that the data are imputed.

The solution, at least with random imputation, is to repeat the imputation process more than once, producing multiple "completed" data sets. Because of the random component, the estimates of the parameters of interest will be slightly different for each imputed data set. This variability across imputations can be used to adjust the standard errors upward.

For the simulated sample of 10,000 cases, I repeated the random imputation process eight times, yielding the estimates in Table 5.1. Although these estimates are approximately unbiased, the standard errors are downwardly biased because they do not take the imputation[9] into account. We can combine the eight correlation estimates into a single estimate simply by taking their mean, which is 0.3125. An improved estimate of the standard error takes three steps.

TABLE 5.1
Correlations and Standard Errors for Randomly Imputed Data

Correlation	*S.E.*
0.3159	0.00900
0.3108	0.00903
0.3135	0.00902
0.3210	0.00897
0.3118	0.00903
0.3022	0.00909
0.3189	0.00898
0.3059	0.00906

1. Square the estimated standard errors (to get variances) and average the results across the eight replications.
2. Calculate the variance of the correlation estimates across the eight replications.
3. Add the results of steps 1 and 2 together (applying a small correction factor to the variance in step 2) and take the square root.

To put this into one formula, let M be the number of replications, let r_k be the correlation in replication k, and let s_k be the estimated standard error in replication k. Then the estimate of the standard error of $\bar{r}$ (the mean of the correlation estimates) is

$$\text{S.E.}(\bar{r}) = \sqrt{\frac{1}{M}\sum_k s_k^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_k (r_k - \bar{r})^2}. \qquad [5.1]$$

This formula can be used for any parameter estimated by multiple imputation, with r_k denoting the kth estimate of the parameter of interest (Rubin, 1987). Applying this formula to the correlation example, we get a standard error of 0.01123, which is about 24% higher than the mean of the standard errors in the eight samples.

Allowing for Random Variation in the Parameter Estimates

Although the method I just described for imputing missing values is pretty good, it is not ideal. To generate the imputations for X, I regressed X on Y for the cases with complete data to produce the regression equation

$$\hat{x}_i = a + by_i.$$

For cases with missing data on X, the imputed values were calculated as

$$\tilde{x}_i = a + by_i + s_{x\cdot y}u_i,$$

where u_i is a random draw from a standard normal distribution and $s_{x\cdot y}$ is the estimated standard deviation of the error term (the root mean squared error). For the simulated data set, we had $a = -0.0015$, $b = 0.3101$, and $s_{x\cdot y} = 0.9525$. These values were used to produce imputations for each of the eight completed data sets.

The problem with this approach is that it treats a, b, and $s_{x\cdot y}$ as though they were the true parameters, not sample estimates. Obviously, we cannot know what the true values are, but for "proper" multiple imputations (Rubin, 1987), each imputed data set should be based on a different set of values of a, b, and $s_{x\cdot y}$. These values should be random draws from the Bayesian posterior distribution of the parameters. Only in this way can multiple imputation completely embody our uncertainty with regard to the unknown parameters.

This claim naturally raises several questions. What is the Bayesian posterior distribution of the parameters? How do we get random draws from the posterior distribution? Do we really need this additional complication? The first question really requires another book and, fortunately, there is a good one in the *Quantitative Applications in the Social Sciences* series (Iversen, 1985). As for the second question, there are several different approaches to getting random draws from the posterior distribution, some of them embodied in easy-to-use software. Later in this chapter, when we consider MI under the multivariate normal model, I will explain one method called data augmentation (Schafer, 1997).

Can you get by without making random draws from the posterior distribution of the parameters? It is important to answer to this question, because some random imputation software—like the missing data module in SPSS—does not randomly draw the parameter values. In many cases, I think the answer is yes. If the sample is large and the proportion of cases with missing data is small, MI without

TABLE 5.2
Correlations and Standard Errors for Randomly Imputed Data Using the Data Augmentation Method

Correlation	*S.E.*
0.30636	0.0090614
0.31316	0.0090193
0.31837	0.0089864
0.31142	0.0090302
0.32086	0.0089705
0.29760	0.0091143
0.32701	0.0089306
0.30826	0.0090498

this extra step typically will yield results that are very close to those obtained with it. On the other hand, if the sample is small or if the proportion of cases with missing data is large, the additional variation can make a noticeable difference.

Continuing our correlation example, I imputed eight new data sets using the data augmentation method to generate random draws from the posterior distribution of the parameters. Table 5.2 gives the correlation between X and Y, and its standard error for each data set. The mean of the correlation estimates is 0.31288. Using formula 5.1, the estimated standard error is 0.01329, slightly larger than the 0.01123 obtained with the cruder imputation method. In general, the standard errors will be somewhat larger when the parameters used in the imputation are drawn randomly.

Multiple Imputation Under the Multivariate Normal Model

To do multiple imputation, you need a model to generate the imputations. For the two-variable example just considered, I employed a simple regression model with normally distributed errors. Obviously, more complicated situations require more complicated models. As we saw in Chapter 4, maximum likelihood also requires a model. However, MI is probably less sensitive than ML to the choice of model because the model is used only to impute the missing data, not to estimate other parameters.

Ideally, the imputation model would be specially constructed to represent the particular features of each data set. In practice, it is more convenient to work with "off-the-shelf" models that are easy to use and provide reasonably good approximations for a wide range of data sets.

The most popular model for MI is the multivariate normal model, which previously was used in Chapter 4 as the basis for ML estimation of linear models with missing data. The multivariate normal model implies that

- All variables have normal distributions
- Each variable can be represented as a linear function of all the other variables, together with a normal, homoscedastic error term

Although these are strong conditions, in practice the multivariate normal model seems to do a good job of imputation even when some of the variables have distributions that are manifestly not normal (Schafer, 1997). It is a completely innocuous assumption for those variables that have no missing data. For those variables that do have missing data, normalizing transformations can greatly improve the quality of the imputations.

In essence, MI under the multivariate normal model is a generalization of the method used in the two-variable example of the previous section. For each variable with missing data, we estimate the linear regression of that variable on all other variables of interest. Ideally, the regression parameters are random draws from the Bayesian posterior distribution. The estimated regression equations are then used to generate predicted values for the cases with missing data. Finally, to each predicted value, we add a random draw from the residual normal distribution for that variable.

The most complicated part of the imputation process is getting random draws from the posterior distribution of the regression parameters. As of this writing, two algorithms that accomplish this have been implemented in readily available software: data augmentation (Schafer, 1997) and sampling importance/resampling (SIR; Rubin, 1987). Here are some computer programs that implement these methods:

Data Augmentation

NORM A freeware package developed by Schafer and described in his 1997 book. Available in either a stand-alone Windows version or as an S-PLUS library (http://www.stat.psu.edu/~jls/).

SOLAS A stand-alone commercial package that includes both data augmentation (version 2 and later) and a propensity score method. The latter method is invalid for many applications (Allison, 2000) (http://www.statsolusa.com)

PROC MI A SAS procedure available in release 8.1 and later (http://www.sas.com).

Sampling Importance/Resampling

AMELIA A freeware package developed by King, Honaker, Joseph, Scheve, and Singh (1999). Available either as a stand-alone Windows program or as a module for Gauss (http://gking.harvard.edu/stats.shtml).

SIRNORM A SAS macro written by C. H. Brown and X. Ling (http://yates.coph.usf.edu/research/psmg/web.html).

Both algorithms have some theoretical justification. Proponents of SIR (King, Honaker, Joseph, & Scheve, 2001) claim that it requires far less computer time. However, the relative superiority of these two methods is far from settled. Because I have much more experience with data augmentation, I will focus on this method in the remainder of this chapter.

Data Augmentation for the Multivariate Normal Model

Data augmentation (DA) is a type of Markov chain Monte Carlo (MCMC) algorithm, a general method for finding posterior distributions that has become increasingly popular in Bayesian statistics. In this section, I will describe how it works for the multivariate normal model. Although the available software performs most of the operations automatically, it is helpful to have a general idea of what is going on, especially when things go wrong.

The general structure of the iterative algorithm is much like the EM algorithm for the multivariate normal model, described in the last chapter, except that random draws are made at two points, described subsequently. Before beginning DA, it is necessary to choose a set of

variables for the imputation process. Obviously this should include all variables with missing data, as well as other variables in the model to be estimated. It is also worthwhile to include additional variables (not in the intended model) that are highly correlated with the variables that have missing data or that are associated with the probability that those variables have missing data.

Once the variables are chosen, DA consists of the following steps.

1. Choose starting values for the parameters. For the multivariate normal model, the parameters are the means and the covariance matrix. Starting values can be gotten from standard formulas using listwise deletion or pairwise deletion. Even better are the estimates obtained with the EM algorithm described in the last chapter.
2. Use the current values of the means and covariances to obtain estimates of regression coefficients for equations in which each variable with missing data is regressed on all observed variables. This is done for each pattern of missing data.
3. Use the regression estimates to generate predicted values for all the missing values. To each predicted value, add a random draw from the residual normal distribution for that variable.
4. Using the "completed" data set, with both observed and imputed values, recalculate the means and covariance matrix using standard formulas.
5. Based on the newly calculated means and covariances, make a random draw from the posterior distribution of the means and covariances.
6. Using the randomly drawn means and covariances, go back to step 2 and continue cycling through the subsequent steps until convergence is achieved. The imputations that are produced during the final iteration are used to form a completed data set.

Step 5 needs a little more explanation. To get the posterior distribution of the parameters, we first need a "prior" distribution. Although this can be based on prior beliefs about those parameters, the usual practice is to use a "noninformative" prior, that is, a prior distribution that contains little or no information about the parameters. Here is how it works in a simple situation. Suppose we have a sample of size n with measurements on a single, normally distributed variable Y. The sample mean is $\bar{y}$ and the sample variance is s^2. We want random draws from the posterior distribution of μ and σ^2. With a noninformative prior,[10] we can get $\tilde{\sigma}^2$, a random draw for the variance, by sampling from a chi-square distribution with $n-1$ degrees

of freedom, taking the reciprocal of the drawn value, and multiplying the result by ns^2. We then get a random draw for the mean by sampling from a normal distribution with mean $\bar{y}$ and variance $\tilde{\sigma}^2/n$.

If there were no missing data, these would be random draws from the true posterior distribution of the parameters, but if we have imputed any missing data, what we actually have are random draws from the posterior distribution that would result if the imputed data were the true data. Similarly, when we randomly impute missing data in step 3, what we have are random draws from the posterior distribution of the missing data, given the current parameter values. However, because the current values may not be the true values, the imputed data may not be random draws from the true posterior distribution. That is why the procedure must be iterative. By continually moving back and forth between random draws of parameters (conditional on both observed and imputed data) and random draws of the missing data (conditional on the current parameters), we eventually get random draws from the joint posterior distribution of both data and parameters, conditioning only on the observed data.

Convergence in Data Augmentation

When you do data augmentation, you must specify the number of iterations. However, that raises a tough question: How many iterations are necessary to get convergence to the joint posterior distribution of missing data and parameters? With iterative estimation for maximum likelihood, as in the EM algorithm, the estimates converge to a single set of values. Convergence can then be easily assessed by checking to see how much change there is in the parameter estimates from one iteration to the next. For data augmentation, on the other hand, the algorithm converges to a probability distribution, not a single set of values. That makes it rather difficult to determine whether convergence has, in fact, been achieved. Although some diagnostic statistics are available for assessing convergence (Schafer, 1997), they are far from definitive.

In most applications, the choice of number of iterations will be a stab in the dark. To give you some idea of the range of possibilities, Schafer (1997) used anywhere between 50 and 1,000 iterations for the examples in his book. The more the better, but each iteration can be computationally intensive, especially for large samples with lots of

variables. Specifying a large number of iterations can leave you staring at your monitor for painfully long periods of time.

There are a couple of principles to keep in mind. First, the higher is the proportion of missing data (actually, missing *information*, which is not quite the same thing), the more iterations will be needed to reach convergence. If only 5% of the cases have any missing data, you can probably get by with only a small number of iterations. Second, the rate of convergence of the EM algorithm is a useful indication of the rate of convergence for data augmentation. A good rule of thumb is that the number of iterations for DA should be at least as large as the number of iterations required for EM. That is another reason for always running EM before data augmentation. (The first reason is that EM gives good starting values for data augmentation.)

My own feeling about the iteration issue is that it is not all that critical in most applications. Moving from deterministic imputation to randomized imputation is a huge improvement, even if the parameters are not randomly drawn. Moving from randomized imputation without random parameter draws to randomized imputation *with* random parameter draws is another substantial improvement, but not nearly as dramatic. Moving from a few iterations of data augmentation to many iterations improves things still further, but the marginal return is likely to be quite small in most applications.

An additional complication stems from the fact that multiple imputation produces multiple data sets. At least two are required, but the more the better. For a fixed amount of computing time, one can either produce more data sets or more iterations of data augmentation per data set. Unfortunately, data sets with lots of missing information need both more iterations and more data sets. Although little has been written about this issue, I tend to think that more priority should be put on additional data sets.

Sequential Versus Parallel Chains of Data Augmentation

We have just seen how to use data augmentation to produce a single, completed data set. For multiple imputation, we need several data sets. Two methods have been proposed to do this:

1. *Parallel.* Run a separate chain of iterations for each of the desired data sets. These might be from the same set of starting values (say, the EM estimates) or different starting values.

types and missing data patterns. As a routine method for handling missing data, it is probably the best that is currently available. There are, however, several alternative approaches that may be preferable in some circumstances.

One of the most obvious limitations of the multivariate normal model is that it is designed only to impute missing values for quantitative variables. As we have seen, categorical variables can be accommodated by using some ad hoc fixups. However, sometimes you may want to do better. For situations in which *all* variables in the imputation process are categorical, a more attractive model is the unrestricted multinomial model (which has a parameter for every cell in the contingency table) or a log-linear model that allows restrictions on the multinomial parameters. In Chapter 4, we discussed ML estimation of these models. Schafer (1997) showed how these models also can be used as the basis for data augmentation to produce multiple imputations and he developed a freeware program called CAT to implement the method (http://www.stat.psu.edu/~jls/).

Another Schafer program (MIX) uses data augmentation to generate imputations when the data consist of a mixture of categorical and quantitative variables. This method presumes that the categorical variables have a multinomial distribution, possibly with log-linear restrictions on the parameters. Within each cell of the contingency table created by the categorical variables, the quantitative variables are assumed to have a multivariate normal distribution. The means of these variables are allowed to vary across cells, but the covariance matrix is assumed to be constant.

At this writing, both CAT and MIX are available only as libraries to the S-PLUS statistical package, although stand-alone versions are promised. In both cases, the underlying models potentially have many more parameters than the multivariate normal model. As a result, effective use of these methods typically requires more knowledge and input from the person performing the imputation, together with larger sample sizes to achieve stable estimates.

If data are missing for a single categorical variable, multiple imputation under a logistic (logit) regression model is reasonably straightforward (Rubin, 1987). Suppose data are missing on marital status, coded into five categories, and there are several potential predictor variables, both continuous and categorical. For the purposes of imputation, we estimate a multinomial logit model for marital status as a function of the predictors, using cases with complete data. This produces a set of

coefficient estimates $\hat{\beta}$ and an estimate of the covariance matrix $\widehat{V}(\hat{\beta})$. To allow for variability in the parameter estimates, we take a random draw from a normal distribution with a mean of $\hat{\beta}$ and a covariance matrix $\widehat{V}(\hat{\beta})$. (Schafer [1997] gave practical suggestions on how to do this efficiently.) For each case with missing data, the drawn coefficient values and the observed covariate values are substituted into the multinomial logit model to generate predicted probabilities of falling into the five marital status categories. Based on these predicted probabilities, we randomly draw one of the marital status categories as the final imputed value.[12] The whole process is repeated multiple times to generate multiple completed data sets. Of course, a binary variable would be just a special case of this method. This approach also can be used with a variety of other parametric models, including Poisson regression and parametric failure-time regressions.

Nonparametric and Partially Parametric Methods

Many methods have been proposed for doing multiple imputation under less stringent assumptions than the fully parametric methods we have just considered. In this section, I will consider a few representative approaches, but keep in mind that each of these approaches has many different variations. All these methods are most naturally applied when there are missing data on only a single variable, although they often can be generalized without difficulty to multiple variables when data are missing in a monotone pattern (described in Chapter 4). See Rubin (1987) for details on monotone generalizations. These methods can sometimes be used when the missing data do *not* follow a monotone pattern, but in such settings they typically lack solid theoretical justification.

When choosing between parametric and nonparametric methods, there is the usual trade-off between bias and sampling variability. Parametric methods tend to have less sampling variability, but they may give biased estimates if the parametric model is not a good approximation to the phenomenon of interest. Nonparametric methods may be less prone to bias under a variety of situations, but the estimates often have more sampling variability.

Hot Deck Methods

The best-known approach to nonparametric imputation is the "hot deck" method, which is frequently used by the U.S. Census Bureau to

produce imputed values for public-use data sets. The basic idea is that we want to impute missing values for a particular variable Y, which may be either quantitative or categorical. We find a set of categorical X variables (with no missing data) that are associated with Y. We form a contingency table based on the X variables. If there are cases with missing Y values within a particular cell of the contingency table, we take one or more of the nonmissing cases in the same cell and use their Y values to impute the missing Y values.

Obviously there are a lot of complications that may arise. The critical question is how do you choose which "donor" values to assign to the cases with missing values? Clearly the choice of donor cases should be randomized somehow to avoid bias. This leads naturally to multiple imputation because any randomized method can be applied more than once to produce different imputed values. The trick is to do the randomization in such a way that all the natural variability is preserved. To accomplish this, Rubin proposed a method he coined the approximate Bayesian bootstrap (Rubin, 1987; Rubin & Schenker, 1991). Here is how it is done. Suppose that in a particular cell of the contingency table there are n_1 cases with complete data on Y and n_0 cases with missing data on Y. Follow these steps:

1. From the set of n_1 cases with complete data, take a random sample (with replacement) of n_1 cases.
2. From this sample, take a random sample (with replacement) of n_0 cases.
3. Assign the n_0 observed values of Y to the n_0 cases with missing data on Y.
4. Repeat steps 1 to 3 for every cell in the contingency table.

These four steps produce one completed data set when applied to all cells of the contingency table. For multiple imputation, the whole process is repeated multiple times. After the desired analysis is performed on each data set, the results are combined using the same formulas we used for multivariate normal imputations.

Although it might seem that we could skip step 1 and directly choose n_0 donor cases from among the n_1 cases with complete data, this does not produce sufficient variability for estimating standard errors. Additional variability comes from the fact that sampling in step 2 is with replacement.

Predictive Mean Matching

A major attraction of hot deck imputation is that the imputed values are all actual observed values. Consequently, there are no "impossible" or out-of-range values, and the shape of the distribution tends to be preserved. A disadvantage is that the predictor variables all must be categorical (or treated as such), which imposes serious limitations on the number of possible predictor variables. To remove this limitation, Little (1988) proposed a partially parametric method called *predictive mean matching*. Like the multivariate normal parametric method, this approach begins by regressing Y, the variable to be imputed, on a set of predictors for cases with complete data. This regression is then used to generate predicted values for both the missing and the nonmissing cases. Then, for each case with missing data, we find a set of cases with complete data that have predicted values of Y that are "close" to the predicted value for the case with missing data. From this set of cases, we randomly choose one case whose Y value is donated to the missing case.

For a single Y variable, it is straightforward to define closeness as the absolute difference between predicted values. However, then we must decide how many of the close predicted values to include in the donor pool for each missing case or, equivalently, what should be the cutoff point in closeness for forming the set of possible donor values? If a small donor pool is chosen, there will be more sampling variability in the estimates. On the other hand, too large a donor pool can lead to possible bias because many donors may be unlike the recipients. To deal with this ambiguity, Schenker and Taylor (1996) developed an "adaptive method" that varies the size of the donor pool for each missing case based on the "density" of complete cases with close predicted values. They found that their method did somewhat better than methods with fixed size donor pools of either 3 or 10 closest cases. However, the differences among the three methods were sufficiently small that the adaptive method hardly seems worth the extra computational cost.

In doing predictive mean matching, it is also important to adjust for the fact that the regression coefficients are only estimates of the true coefficients. As in the parametric case, this can be accomplished by randomly drawing a new set of regression parameters from their posterior distribution before calculating predicted values for each

imputed data set. Here is how to do it:

1. Regress Y on X (a vector of covariates) for the n_1 cases with no missing data on Y, producing regression coefficients b (a $k\times$ 1 vector) and residual variance estimate s^2.
2. Make a random draw from the posterior distribution of the residual variance (assuming a noninformative prior). This is accomplished by calculating $(n_1 - k)\ s^2/\chi^2$, where χ^2 represents a random draw from a chi-square distribution with $n_1 - k$ degrees of freedom. Let $s^2_{[1]}$ be the first such random draw.
3. Make a random draw from the posterior distribution of the regression coefficients. This is accomplished by drawing from a multivariate normal distribution with mean b and covariance matrix $s^2_{[1]}(\mathbf{X}'\mathbf{X})^{-1}$, where $\mathbf{X}$ is an $n_1 \times k$ matrix of X values. Let $b_{[1]}$ be the first such random draw. See Schafer (1997) for practical suggestions on how to do this.

For each new set of regression parameters, predicted values are generated for all cases. Then, for each case with missing data on Y, we form a donor pool based on the predicted values and randomly choose one of the observed values of Y from the donor pool. This approach to predictive mean matching can be generalized to more than one Y variable with missing data, although the computations may become rather complex (Little, 1988).

Sampling on Empirical Residuals

In the data augmentation method, residual values are sampled from a standard normal distribution and then added to the predicted regression values to get the final imputed values. We can modify this method to be less dependent on parametric assumptions by making random draws from the actual set of residuals produced by the linear regression. This can yield imputed values whose distribution is more like that of the observed variable (Rubin, 1987), although it is still possible to get imputed values that are outside the permissible range.

As with other approaches to multiple imputation, there are some important subtleties involved in doing this properly. As before, let Y be the variable with missing data to be imputed for n_0 cases, with observed data on n_1 cases. Let X be a $k\times1$ vector of variables (including a constant) with no missing data on the n_1 cases. We begin by performing the preceding three steps to obtain the linear regression of

Y on X and generate random draws from the posterior distribution of the parameters. Then we add the following steps:

4. Based on the regression estimates in step 1, calculate standardized residuals for the cases with no missing data:

$$e_i = (y_i - bx_i)/\sqrt{s^2(1 - k/n_1)}$$

5. Draw a simple random sample (with replacement) of n_0 values from the n_1 residuals calculated in step 4.
6. For the n_0 cases with missing data, calculate imputed values of Y as

$$y_i = b_{[1]}x_i + s_{[1]}e_i,$$

where e_i represents the residuals drawn in step 5, and $b_{[1]}$ and $s_{[1]}$ are the first random draws from the posterior distribution of the parameters.

These six steps produce one completed set of data. To get additional data sets, simply repeat steps 2 through 6 (except for step 4, which should not be repeated).

As Rubin (1987) explained, this methodology can be readily extended to data sets with a monotonic missing pattern on several variables. Each variable is imputed using as predictors all variables that are observed when it is missing. The empirical residual method also can be modified to allow for heteroscedasticity in the imputed values (Schenker & Taylor, 1996). For each case to be imputed, the pool of residuals is restricted to those observed cases that have predicted values of Y that are close to the predicted value for the case with missing data.

Example

Let us try the partially parametric methods on a subset of the college data. TUITION is fully observed for 1,272 colleges. (For simplicity, we shall exclude the 30 cases with missing data on this variable.) Of these 1,272 colleges, only 796 report BOARD, the annual average cost of board at each college. Using TUITION as a predictor, our goal is to impute the missing values of BOARD for the other 476 colleges, and estimate the mean of BOARD for all 1,272 colleges.

First, we apply the methods we have used before. For the 796 colleges with complete data (listwise deletion), the average BOARD is \$2,060 with a standard error of 23.4. Applying the EM algorithm to

TUITION and BOARD, we get a mean BOARD of 2,032 (but no standard error). The EM estimate of the correlation between BOARD and TUITION was 0.555. Multiple imputation under the multivariate normal model using data augmentation gave a mean BOARD of 2,040 with an estimated standard error of 21.2.

Because BOARD is highly skewed to the right, there is reason to suspect that the multivariate normal model may not be appropriate. Quite a few of the values imputed by data augmentation were less than the minimum observed value of 531, and one imputed value was negative. Perhaps we can do better by sampling on the empirical residuals. For the 796 cases with data on both TUITION and BOARD, the ordinary least-squares (OLS) regression of BOARD on TUITION was

$$\text{BOARD} = 1{,}497.4 + 67.65 * \text{TUITION}/1{,}000$$

with a root mean squared error (rmse) estimated at 542.6. Standardized residuals from this regression were calculated for the 796 cases.

The estimated regression parameters were used to make five random draws from the posterior distribution of the parameters as in steps 2 and 3 (assuming a noninformative prior). The drawn values were

Intercept	*Slope*	*rmse*
1536.40	66.6509	531.990
1503.65	71.5916	552.708
1501.61	66.9756	554.800
1486.84	66.9850	548.400
1504.23	61.2308	534.895

To create the first completed data set, 476 residual values were randomly drawn with replacement from among the 796 cases. These standardized residuals were assigned arbitrarily to the 476 cases with missing data on BOARD. Letting E be the assigned residual for a given case, the imputed values for BOARD were generated as

$$\text{BOARD} = 1{,}536.40 + 66.6509 * \text{TUITION}/1{,}000 + 531.990 * E.$$

This process was repeated for the four remaining data sets, with new sampling on the residuals and new values of the regression parameters at each step.

Once the five data sets were produced, the mean and standard error were computed for each data set, and the results were combined using formula 5.1 for the standard error. The final estimate for the mean of BOARD was 2,035 with an estimated standard error of 20.4, which is quite close to multiple imputation based on a normal distribution.

Now let us try predictive mean matching. Based on the coefficients from the OLS regression of BOARD on TUITION, I generated five new random draws from the posterior distribution of the regression parameters:

Intercept	*Slope*	*rmse*
1465.89	67.8732	557.531
1548.98	64.5723	539.952
1428.82	67.3901	512.381
1469.34	67.3750	550.945
1517.92	66.1926	534.804

For the first set of parameter values, I generated predicted values of BOARD for all cases, both observed and missing. For each case with missing data on BOARD, I found the five observed cases whose predicted values were closest to the predicted value for the case with missing data. I randomly chose one of those five cases and assigned its *observed* value of BOARD as the imputed value for the missing case. This process was repeated for each of the five sets of parameter values to produce five complete data sets. (It is just coincidence that the number of data sets is the same as the number of observed cases matched to each missing case.) The mean and standard error were then computed for each data set and the results were combined in the usual way. The combined mean of BOARD was 2,028 with an estimated standard error of 23.0.

All four imputation methods produced similar estimates of the means and all were noticeably lower than the mean based on listwise deletion. Schenker and Taylor (1996) suggested that although parametric and partially parametric imputation methods tend to yield quite similar estimates of mean structures (including regression coefficients), they may produce more divergent results for the marginal distribution of the imputed variables. Their simulations indicated that for applications where the marginal distribution is of major interest, partially parametric models have a distinct advantage. This was especially true when the regressions used to generate predicted values were misspecified in various ways.

Sequential Generalized Regression Models

One of the attractions of data augmentation is that, unlike the nonparametric and semiparametric methods just discussed, it easily can handle data sets with a substantial number of variables with missing data. Unfortunately, this method requires specifying a multivariate distribution for all the variables, and that is not an easy thing to do when the variables are of many different types, for example, continuous, binary, and count data. Another approach has been proposed for handling missing data in large, complex data sets with several different variable types. Instead of fitting a single comprehensive model (e.g., the multivariate normal), a separate regression model is specified for each variable that has any missing data. For each dependent variable, the regression model is chosen to reflect the type of data. The method involves cycling through several regression models, imputing missing values at each step.

Although this approach is very appealing, it does not yet have as strong a theoretical justification as the other methods we have considered. At this writing, the only detailed accounts are the unpublished reports of Brand (1999), Van Buuren and Oudshoorn (1999), and Raghunathan, Lepkowski, Van Hoewyk, and Solenberger (1999). In the Raghunathan et al. version of this method, the available models include normal linear regression, binary logistic regression, multinomial logit, and Poisson regression. The regression models are estimated in a particular order, beginning with the dependent variable with the least missing data and proceeding to the dependent variable with the most missing data. Let us denote these variables by Y_1 through Y_k and let X denote the set of variables with no missing data.

The first "round" of estimation proceeds as follows. Regress Y_1 on X and generate imputed values using a method similar to that described previously for the multinomial logit model in the section "Other Parametric Approaches to Multiple Imputation." Bounds and restrictions may be placed on the imputed values. Then regress Y_2 on X and Y_1, including the imputed values of Y_1, and generate imputed values for Y_2. Then regress Y_3 on X, Y_1, and Y_2 (including imputed values on both Ys). Continue until all the regressions have been estimated. The second and subsequent rounds repeat this process, except that now each variable is regressed on *all* other variables using any imputed values from previous steps. The process continues for a prespecified number of rounds or until stable imputed values

occur. A SAS macro for accomplishing these tasks is available at http://www.isr.umich.edu/src/smp/ive.

For their version of the method, Van Buuren and Oudshoorn coined the name MICE (for multiple imputation by chained equations), and they developed S-PLUS functions to implement it (available at http://www.multiple-imputation.com/). The major differences between their approach and that of Raghunathan et al. is that MICE does not include Poisson regression, but does allow more options (both parametric and partially parametric) in the methods for random draws of imputed values.

Linear Hypothesis Tests and Likelihood Ratio Tests

To this point, our approach to statistical inference with multiple imputation has been very simple. For a given parameter, the standard error of the estimate is calculated using formula 5.1. This standard error is then plugged into conventional formulas based on the normal approximation to produce a confidence interval or a t statistic for some hypothesis of interest. Sometimes this is not enough. Often we want to test hypotheses about sets of parameters, for example, that two parameters are equal to each other or that several parameters are all equal to zero. These sorts of hypotheses are particularly relevant when we estimate several coefficients for a set of dummy variables. In addition, there is often a need to compute likelihood ratio statistics by comparing one model with another, simpler model. Accomplishing these tasks is not so straightforward when doing multiple imputation. Schafer (1997) described three different approaches, none of which is totally satisfactory. I will briefly describe them here and we will look at an example in the next section.

Wald Tests Using Combined Covariance Matrices

When there are no missing data, a common approach to multiple parameter inference is to compute Wald chi-square statistics based on the parameter estimates and their estimated covariance matrix. Here is a review, which, unfortunately, requires matrix algebra. Suppose we want to estimate a $p \times 1$ parameter vector $\boldsymbol{\beta}$. We have estimates $\hat{\boldsymbol{\beta}}$ and estimated covariance matrix $\mathbf{C}$. We want to test a linear hypothesis expressed as $\mathbf{L}\boldsymbol{\beta} = \mathbf{c}$, where $\mathbf{L}$ is an $r \times p$ matrix of constants and $\mathbf{c}$ is an $r \times 1$ vector of constants. For example, if we want to test the

hypothesis that the first two elements of $\boldsymbol{\beta}$ are equal to each other, we need $\mathbf{L} = [1 \quad -1 \quad 0 \quad 0 \quad 0 \quad \cdots \quad 0]$ and $\mathbf{c} = 0$. The Wald test is computed as

$$W = (\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{LCL}']^{-1}(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}), \qquad [6.1]$$

which has an approximate chi-square distribution with r degrees of freedom under the null hypothesis.[13]

Now we generalize this method to the multiple imputation setting. Instead of $\hat{\boldsymbol{\beta}}$, we can use $\bar{\boldsymbol{\beta}}$, the mean of the estimates across several completed data sets, that is,

$$\bar{\boldsymbol{\beta}} = \frac{1}{M-1}\sum_k \hat{\boldsymbol{\beta}}_k.$$

Next we need an estimate of the covariance matrix that combines the within-sample variability and the between-sample variability. Let $\mathbf{C}_k$ be the estimated covariance matrix for the parameters in data set k and let $\overline{\mathbf{C}}$ be the average of those matrices across the M data sets. The between-sample variability is defined as

$$\mathbf{B} = \frac{1}{M-1}\sum_k (\hat{\boldsymbol{\beta}}_k - \bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_k - \bar{\boldsymbol{\beta}})'.$$

The combined estimate of the covariance matrix is then

$$\widetilde{\mathbf{C}} = \overline{\mathbf{C}} + (1 + 1/M)\mathbf{B},$$

which is just a multivariate generalization of formula 5.1 without the square root. We get our test statistic by formula 6.1 with $\bar{\boldsymbol{\beta}}$ and $\widetilde{\mathbf{C}}$ substituted for $\hat{\boldsymbol{\beta}}$ and $\mathbf{C}$.

Unfortunately, this does not work well in the typical case where M is 5 or less. In such cases, $\mathbf{B}$ is a rather unstable estimate of the between-sample covariance, and the resulting distribution of W is not chi-square. Schafer (1997) gave a more stable estimator for the covariance matrix, but this required the unreasonable assumption that the fraction of missing information is the same for all the elements of $\hat{\boldsymbol{\beta}}$. Nevertheless, some simulations show that this alternate method works well even when the assumption is violated. This method has been incorporated into the SAS procedure MIANALYZE.

Likelihood Ratio Tests

If the model of interest is estimated by maximum likelihood and there are no missing data, multiparameter tests are often performed by computing likelihood ratio chi-squares. The procedure is quite simple. Let l_0 be the log-likelihood for a model that imposes the hypothesis and let l_1 be the log-likelihood for a model that relaxes the hypothesis. The likelihood ratio statistic is just $L = 2(l_1 - l_0)$.

As before, our goal is to generalize this to multiple imputation. The first step is to perform the desired likelihood ratio test in each of the M completed data sets. Let $\overline{L}$ be the mean of the likelihood ratio chi-squares computed across those M data sets. That is the easy part. Now comes the hard part. To get those chi-squares, it was necessary to estimate two models in each data set, one with the hypothesis imposed and one with the hypothesis relaxed. Let $\bar{\boldsymbol{\beta}}_0$ be the mean of the M parameter estimates when the hypothesis is imposed and let $\bar{\boldsymbol{\beta}}_1$ be the mean of the parameter estimates when the hypothesis is relaxed. In each data set, we then compute the log-likelihood for a model with parameter values forced to be $\bar{\boldsymbol{\beta}}_0$ and again for a model with parameters set at $\bar{\boldsymbol{\beta}}_1$. (This obviously requires that the software be able to calculate and report log-likelihoods for user-specified parameter values.) Based on these two log-likelihoods, a likelihood ratio chi-square is computed in each data set. Let $\widetilde{L}$ be the mean of these chi-square statistics across the M samples.

The final test statistic is then $\widetilde{L}/(r + (\frac{M+1}{M-1})(\overline{L} - \widetilde{L}))$, where r is the number of restrictions imposed by the hypothesis. Under the null hypothesis, this statistic has approximately an F distribution with numerator degrees of freedom equal to r. The denominator degrees of freedom (d.d.f) is somewhat awkward to calculate. Let $t = r(M-1)$ and let

$$q = \left(\frac{M+1}{M-1}\right)\left(\frac{\overline{L} - \widetilde{L}}{r}\right).$$

If $t > 4$, the d.d.f. $= 4 + (t-4)[1 + (1 - 2/t)/q]^2$. If $t \leq 4$, the d.d.f. $= t(1 + 1/r)(1 + 1/q)^2/2$.

Combining Chi-Square Statistics

Both the Wald test and the likelihood ratio test lack the appealing simplicity of the single-parameter methods used earlier. In particular,

they require that the analysis software have specialized options and output, something we have generally tried to avoid. I now discuss a third method that is easy to compute from standard output, but may not be as accurate as the other two methods (Li, Meng, Raghunathan, & Rubin, 1991). All that is needed is the conventional chi-square statistic (either Wald or likelihood ratio) calculated in each of the M completed data sets, and the associated degrees of freedom.

Let d_k^2 be a chi-square statistic with r degrees of freedom calculated in data set k. Let $\bar{d}^2$ be the mean of these statistics over the M data sets and let s_d^2 be the sample variance of the *square roots* of the chi-square statistics over the M data sets, that is,

$$s_d^2 = \frac{1}{M-1}\sum_k (d_k - \bar{d})^2.$$

The proposed test statistic is

$$D = \frac{\bar{d}^2/r - (1-1/M)s_d^2}{1+(1+1/M)s_d^2}.$$

Under the null hypothesis, this statistic has approximately an F distribution with r as the numerator degrees of freedom. The denominator degrees of freedom is approximated by

$$\left(\frac{M-1}{r^{3/M}}\right)\left(1+\frac{M}{(M+1/M)s_d^2}\right)^2.$$

I have written a SAS macro (COMBCHI) to perform these computations and compute a p value. It is available on my Web site (http://www.ssc.upenn.edu/~allison). To use it, all you need to do is enter several chi-square values and the degrees of freedom. The macro returns a p value.

MI Example 2

Let us consider another detailed empirical example that illustrates some of the techniques discussed in this chapter. The data set consists of 2,992 respondents to the 1994 General Social Survey (Davis & Smith, 1997). Our dependent variable is SPANKING, a response to

the question, "Do you strongly agree, agree, disagree, or strongly disagree that it is sometimes necessary to discipline a child with a good, hard spanking?" As the question itself indicates, there were four possible ordered responses, coded as integers 1 through 4. By design, this question was part of a module that was administered only to a random two-thirds of the sample. Thus, there were 1,015 cases that were missing completely at random. In addition, another 27 respondents were missing with responses coded "don't know" or "no answer."

Our goal is to estimate an ordered logistic (cumulative logit) model (McCullagh, 1980) in which SPANKING is predicted by the following variables:

AGE	Respondent's age in years, ranging from 18 to 89. Missing 6 cases.
EDUC	Number of years of schooling. Missing 7 cases.
INCOME	Household income, coded as the midpoint of 21 interval categories, in thousands of dollars. Missing 356 cases.
FEMALE	1 = female; 0 = male.
BLACK	1 = black; 0 = white, other.
MARITAL	Five categories of marital status. Missing 1 case.
REGION	Nine categories of region.
NOCHILD	1 = no children; otherwise 0. Missing 9 cases.

One additional variable, NODOUBT, requires further explanation. Respondents were asked about their beliefs in God. There were six response categories ranging from "I don't believe in God" to "I know God really exists and I have no doubts about it." The latter statement was the modal response with 62% of the respondents. However, like the spanking question, this question was part of a module that was only asked of a random subset of 1,386 respondents. So there were 1,606 cases missing by design. Another 60 cases were treated as missing because they said "don't know" or "no answer." As used here, the variable was coded 1 if the respondent had "no doubts"; otherwise it was coded 0.

Most of the missing data are on three variables, SPANKING, NODOUBT, and INCOME. There were five major missing data patterns in the sample, accounting for 96% of respondents:

771 cases	No missing data on any variables
927 cases	Missing NODOUBT only