Educational Research and Evaluation
2001, Vol. 7, No. 4, pp. 353–383

1380-3611/01/0704-353$16.00

A Review of Methods for Missing Data

Therese D. Pigott
Loyola University Chicago, Wilmette, IL, USA

ABSTRACT

This paper reviews methods for handling missing data in a research study. Many researchers use ad hoc methods such as complete case analysis, available case analysis (pairwise deletion), or single-value imputation. Though these methods are easily implemented, they require assumptions about the data that rarely hold in practice. Model-based methods such as maximum likelihood using the EM algorithm and multiple imputation hold more promise for dealing with difficulties caused by missing data. While model-based methods require specialized computer programs and assumptions about the nature of the missing data, these methods are appropriate for a wider range of situations than the more commonly used ad hoc methods. The paper provides an illustration of the methods using data from an intervention study designed to increase students' ability to control their asthma symptoms.

All researchers have faced the problem of missing quantitative data at some point in their work. Research informants may refuse or forget to answer a survey question, files are lost, or data are not recorded properly. Given the expense of collecting data, we cannot afford to start over or to wait until we have developed foolproof methods of gathering information, an unattainable goal. We find ourselves left with the decision of how to analyze data when we do not have complete information from all informants. We are not alone in this problem; the United States Census Bureau has been involved in a debate with the U.S. Congress and the U.S. Supreme Court over the handling of the undercount in the 2000 U.S. Census. Given that most researchers do not have the resources of the U.S. Census Bureau, what are the options available for analyzing data with missing information?

Address correspondence to: Therese D. Pigott, Loyola University Chicago, 1041 Ridge Road, Wilmette, Illinois 60091, USA. Tel: (847) 853-3301. Fax: (847) 853-3375.
E-mail: tpigott@luc.edu

Manuscript submitted: February, 2000
Accepted for publication: November, 2000

The most common method – and the easiest to apply – is the use of only those cases with complete information. Researchers either consciously or by default in a statistical analysis drop informants who do not have complete data on the variables of interest. As an alternative to complete-case analysis, researchers may fill in a plausible value for the missing observations, such as using the mean of the observed cases on that variable. More recently, statisticians have advocated methods that are based on distributional models for the data (such as maximum likelihood and multiple imputation). Much has been published in the statistical literature on missing data (Little, 1992; Little & Rubin, 1987; Schafer, 1997). However, social science researchers have not used these methods nor have they heeded the advice from this work. Using the typical stages of a research study as an organizer, I will provide an overview of the literature on missing data and suggest ways that researchers without extensive statistical backgrounds can handle missing data. I will argue that all researchers need to exercise caution when faced with missing data. Methods for analyzing missing data require assumptions about the nature of the data and about the reasons for the missing observations that are often not acknowledged. When researchers use missing data methods without carefully considering the assumptions required of that method, they run the risk of obtaining biased and misleading results. Reviewing the stages of data collection, data preparation, data analysis, and interpretation of results will highlight the issues that researchers must consider in making a decision about how to handle missing data in their work. The paper focuses on commonly used missing data methods: complete-cases, available-cases, single-value imputation, and more recent model-based methods, maximum likelihood for multivariate normal data, and multiple imputation.

DATA COLLECTION

Avoiding missing data is the optimal means for handling incomplete observations. All experienced researchers take great care in research procedures, in recruiting informants, and in developing measures. Hard as we try, however, most researchers still encounter missing information that may occur for reasons we have not anticipated. During the data collection phase, the researcher has the opportunity to make decisions about what data to collect, and how to monitor data collection. The scale and distribution of the variables in the data and the reasons for missing data are two critical issues for applying the appropriate missing data techniques.

An illustration of these ideas comes from a study of an asthma education intervention in a set of inner-city middle schools (Velsor-Friedrich, in preparation). In each of eight schools, a randomly chosen set of students with asthma participated in an education program designed to increase their knowledge and confidence in controlling their asthma. A set of students also suffering from asthma served as the control group. Two weeks after the intervention, students completed a scale to measure their self-efficacy beliefs with regard to their asthma, and also completed a questionnaire rating the severity of their symptoms over the 2-week period post-treatment. The next two sections focus on the importance for the reasons for missing data, and for the distribution of the variables in the data set in choosing a method for handling missing data.

Reasons for Missing Data

During data collection, the researcher has the opportunity to observe the possible explanations for missing data, evidence that will help guide the decision about what missing data method is appropriate for the analysis. Missing data strategies from complete-case analysis to model-based methods each carry assumptions about the nature of the mechanism that causes the missing data. In the asthma study, several students have missing data on their rating of symptom severity as is expected with students aged 8 to 14. One possible explanation is that students simply forgot to visit the school clinic to fill out the form. If students are missing their symptom severity rating in a random way – because they forgot or for some other reasons not related to their health, the observations from the rest of the students should be representative of the original treatment and control group ratings. Rubin (1976) introduced the term "missing completely at random" (MCAR) to describe data where the complete cases are a random sample of the originally identified set of cases. Since the complete cases are representative of the originally identified sample, inferences based on only the complete cases are applicable to the larger sample and the target population. Complete-case analysis for MCAR data provides results that are generalizable to the target population with one caveat – the estimates will be less precise than initially planned by the researcher since a smaller number of cases are used for estimation.

Another plausible explanation for missing values of symptom severity may relate directly to the missing value. For example, students who miss school because of the severity of their asthma symptoms also will fail to complete the symptom severity rating. The value of the missing variable is directly related to the value of that variable – students suffering severe asthma attacks (high ratings for symptom severity) may be more likely to be missing a value for symptom severity, an example of nonignorable missing data.

With nonignorable missing data, the reasons for the missing observations depend on the values of those variables. In the asthma data, a censoring mechanism may operate where students in the upper tail of the distribution (with high severity of symptoms) are more likely to have missing observations. The optimal time to investigate the possibility of nonignorable missing data on symptom severity is during data collection when we are in the field monitoring data collection. When we suspect a nonignorable missing data mechanism, we need to use procedures much more complex than will be described here. Little and Rubin (1987) and Schafer (1997) discuss methods that can be used for nonignorable missing data. Ruling out a nonignorable response mechanism can simplify the analysis considerably.

A third possibility also exists for the reasons why symptom severity data are missing. For example, younger children may be missing ratings of symptom severity because they have a harder time interpreting the rating form. Younger students' lack of experience or reading skill may lead to a greater chance of missing this variable. Missing values are not missing because these students have severe symptoms (a nonignorable response mechanism), nor are they missing in a way that creates a random sample of responses (MCAR data). Missing values are missing for reasons related to another variable, Age, that is completely observed. Those with smaller values of Age (younger children) tend to be missing symptom severity, regardless of those children's value for symptom severity. Rubin (1976) uses the term missing at random (MAR) to describe data that are missing for reasons related to completely observed variables in the data set.

When data are MCAR or MAR, the response mechanism is termed ignorable. Ignorable response mechanisms are important because when they occur, a researcher can ignore the reasons for missing data in the analysis of the data, and thus simplify the model-based methods used for missing data analysis. (A more thorough discussion of this issue is given by Heitjan & Basu, 1996). Both maximum likelihood and multiple imputation methods require the assumption of an ignorable response mechanism. As discussed later in the paper, it is difficult to obtain empirical evidence about whether or not the data are MCAR or MAR. Recording reasons for missing data can allow the researcher to present a justification for the missing data method used.

One strategy for increasing the probability of an ignorable response mechanism is to use more than one method for collecting important information. Sensitive survey items such as income may produce much missing data, but less sensitive, surrogate variables such as years of education or type of employment may be less subject to missingness. The statistical relationship between income and other income-related variables increases the chance that information lost in missing variables is supplemented by other completely observed variables. Model-based methods use the multivariate relationship between variables to handle the missing data. Thus, the more informative the data set – the more measures we have on important constructs the better the estimation using model-based methods.

Scale and Distribution of Variables

Another issue related to the data collection stage concerns assumptions we make about the distribution of the variables in the model. When considering a statistical model for a study, we choose analysis procedures appropriate to the scale and distribution of the variables. In the model-based methods I will discuss here, the researcher must make the assumption that the data are multivariate normal, that the joint distribution of all variables in the data set (including outcome measures) is a multivariate normal. This assumption at the outset seems to preclude the use of nominal (non-ordered categorical) variables. As Schafer (1997) discusses, this assumption can be relaxed to the assumption that the data are multivariate normal conditional on the fully observed nominal variables. For example, if we gather information on gender and group assignment in a two-group experiment, we will assume that the variables in the data are multivariate normal within each cell defined by the crossing of gender and group (males and females in the treatment and control group). Two implications arise from this assumption. First, the use of the model-based methods that I will describe here requires that the categorical variables in the model are completely observed. As just discussed, one strategy to help ensure completely observed categorical variables is to gather more than one measure of important variables. Second, if categorical variables in the data have high rates of missing observations, then methods using the multivariate normal assumption should not be used. When categorical variables have small amounts of missing values or are completely observed, Schafer (1997) reports on simulation studies that provide evidence of the robustness of the method to moderate departures from normality. In the analysis section, I will return to the implications of assuming multivariate normal data.

During data collection, the researcher has the opportunity to observe reasons for missing data, and to collect more information for variables particularly susceptible to missing values. Complete-case analysis and the model-based methods described here provide trustworthy results only when the assumptions for the response mechanism and distribution of the data hold. As we will see later in the paper for the illustration case, it is too late in the data analysis stage to gather any information about possible reasons for missing data. The next section discusses what we can learn about missing data from the next stage in a research study.

DATA PROCESSING

During the data collection phase, we carefully obtain as much information as possible, trying to get complete data on all informants, and using more than one way to obtain important variables such as income. The next stage involves processing the data, and the critical task for the researcher is to understand the amount and pattern of missing observations. The researcher needs to have an idea of what variables are missing observations to understand the scope of the missing data problem. Typical univariate statistics often do not give a full account of the missing data; researchers also need to understand the amount of data missing about relationships between variables in the data. I will use data from the asthma study (Velsor-Friedrich, in preparation) to illustrate issues that arise at this stage of a research study.

When first processing the data, we often look at univariate statistics such as the mean, standard deviation, and frequencies to check the amount of missing data. Table 1 describes a set of variables from a study examining the effects of a program to increase students' knowledge of their asthma. I am interested in examining how a measure of a student's self-efficacy beliefs about controlling their asthma symptoms relates to a number of predictors. These predictors are Group, participation in a treatment or control group; Docvis, the number of doctor visits in a specified period post-treatment; Symsev, rating of the severity of asthma symptoms post-treatment; Reading, score on state-wide assessment of reading; Age in years; Gender; and Allergy, the number of allergies suffered by the student.

Table 1. Variable Descriptions.

Variable	Definition	Possible values	*M*	(*SD*)	*N*
Asthma belief Survey	Level of confidence in controlling asthma	Range from 1, little confidence to 5, lots of confidence	4.057	(0.713)	154
Group	Treatment or control group	0 = Treatment 1 = Control	0.558	(0.498)	154
Symsev	Severity of asthma symptoms in 2 week period post-treatment	0 = no symptoms 1 = mild symptoms 2 = moderate symptoms 3 = severe symptoms	0.235	(0.370)	141
Reading	Standardized state reading test score	Grade equivalent scores, ranging from 1.10 to 8.10	3.443	(1.636)	79
Age	Age of child in years	Range from 8 to 14	10.586	(1.605)	152
Gender	Gender of child	0 = Male 1 = Female	0.442	(0.498)	154
Allergy	Number of allergies reported	Range from 0 to 7	2.783	(1.919)	83

As seen in Table 1, missing data occurs on six of the seven predictors with Reading and Allergy missing almost half of the observations. Using these statistics may imply that we have complete data on about half of the data set. However, Table 2 presents an alternative data summary, the patterns of missing data that exist. The column totals of Table 2 provide the number of cases missing each variable, similar to the number of values observed as in Table 1. A display similar to the one given here can be generated in the SPSS Missing Value Analysis module (SPSS, 1999) as well as in Schafer's (1999) NORM freeware program. The frequencies of each missing data pattern, given in the row totals, shows that 19 (12.3%) of all cases observe all variables. One hundred and eleven cases (72.1%) are missing just one variable, either Reading or Allergy. The remaining 24 cases (15.5%) are missing two or more variables. It is difficult from the univariate statistics alone to anticipate that only 19 cases have complete data on all variables.

Before selecting an appropriate method for dealing with the missing data problem, we need to make a judgment about the most plausible assumptions for the response mechanism, the reasons for the missing data. As Schafer (1997) discusses, a researcher rarely has detailed information about the

Table 2. Missing Data Patterns.

Symsev	Reading	Age	Allergy	# of cases	% of cases
O	O	O	O	19	12.3
M	O	O	O	1	0.6
O	M	O	O	54	35.1
O	O	O	M	56	36.4
M	M	O	O	9	5.8
M	O	O	M	1	0.6
O	M	O	M	10	6.5
O	O	M	M	2	1.3
M	M	O	M	2	1.3
# missing 13 (8.4%)	# missing 75 (48.7%)	# missing 2 (1.3%)	# missing 71 (46.1)	154	

reasons for missing data. We rely on our knowledge of the data collection procedures as described in the previous section, and our substantive knowledge of the research area. We can use Rubin's (1976) categories to develop conjectures about the reasons for missing data in the asthma data set. Are the data likely missing completely at random (MCAR)? We might eliminate this possibility given our knowledge of the study and the study participants respondents often do not answer questions, especially pre-adolescents and adolescents. Other researchers have suggested empirical ways for examining MCAR. Cohen and Cohen (1975) have suggested developing missing data dummy codes for each variable with missing data, and using this missing variable code as a predictor in a regression model. For example, I could create a variable that takes the value 1 when Allergy is observed and 0 when Allergy is missing. The missing Allergy dummy code could then serve as a predictor in the model, thus allowing the use of all cases. When the regression coefficient for the missing data code is significant, the researcher may infer that the cases missing the variable tend to have a conditional mean value of the outcome different from cases observing that variable. Jones (1996), however, examines the use of missing-indicator variables, finding that this method results in the overestimation of the residual variance of the regression. Alternatively, Little (1988) provides a likelihood ratio test of the assumption of missing completely at random (MCAR). This test is part of the program BMDPAM, in the BMDP (Dixon, 1992) statistical package. In this example, the value of the likelihood ratio test is $\chi^2 = 256.61\ p = 0.000$, indicating that the data in this example are not missing completely at random.

If we reject the MCAR assumption, can we assume the incomplete predictors are missing at random? We do not and cannot have direct information on this assumption. Assuming that the data are MAR means that the reasons for missing values on five of the predictor variables depend on one or more of the completely observed variables in the data set: Asthma Belief Score, Group, Age and Gender. As in the MCAR case, we cannot gather direct evidence about whether MAR is a reasonable assumption for the data.

Alternatively, we could assume that the variables with missing values are subject to a censoring mechanism where the probability of observing a value depends on the actual value of that variable. For example, we may be more likely to observe a Reading score in the middle or upper part of the distribution rather than at the lower tail. When the researcher assumes missing data result from a nonignorable response mechanism, more complex methods of analysis apply, namely adding an explicit model for the response mechanism to the model. Schafer (1997) discusses options when the missing data mechanism is nonignorable.

Given that one cannot obtain direct, empirical information about the response mechanism, an alternative strategy is to examine the sensitivity of results to the MCAR and MAR assumptions by comparing several analyses, such as complete case and the model-based methods described in the next section. Differences in results across several analyses may provide some information about what assumptions may be most relevant.

The first two stages of a research study provide the researcher with information about the nature of the data – the number and pattern of missing observations, plausible reasons for the missing data, and descriptive statistics of the observed data. Having this information may help a researcher to choose an appropriate analysis, or conduct a sensitivity analysis when missing data occur.

DATA ANALYSIS

Many researchers are limited in their data analysis options by the statistical computing packages available, as well as knowledge of alternatives. The most widely used methods are complete case analysis, available case analysis, and single-value imputation. Another set of strategies, maximum likelihood and multiple imputation, are based on models for the data, in this case, the multivariate normal distribution. This section focuses on these five methods for analyzing data with missing observations, examining the assumptions of each method, and describing how one analyzes the data with current software. Although current software provides options such as complete case analysis, available case analysis, and mean substitution, these methods do not have equivalent assumptions about the data, nor do they provide similar results. Many statistical techniques appeal to researchers because of their applicability in a wide range of contexts and data given a set of clearly defined assumptions. Only complete case analysis and model-based methods can apply broadly across a number of data contexts with the fewest number of assumptions.

While the discussion thus far has not distinguished between problems with missing outcomes and missing predictor variables in an analysis, the researcher does need to keep in mind the nature of the variables with missing observations, and the goal of the statistical analysis. As Little (1992) discusses, missing outcomes pose different problems from missing predictors. When the outcomes are MAR, those cases with missing outcomes and completely observed predictors do not contribute any information to a linear model looking at the relationships between the outcome and the predictors. When outcomes and predictors are MAR, cases missing both outcomes and predictors contribute some information about the joint distribution of the missing predictors. When outcomes are not MAR, then more complex modeling procedures are needed. The methods described here will focus on MAR data only.

Commonly-Used Missing Data Methods

Complete-Case Analysis

When a researcher is estimating a model, such as a linear regression, most statistical packages use listwise deletion by default. Cases that are missing variables in the proposed model are dropped from the analysis, leaving only complete cases. A researcher using complete cases assumes that the observed complete cases are a random sample of the originally targeted sample, or in Rubin's (1976) terminology, that the missing data are MCAR. When a data set has only a few missing observations, the assumption of MCAR data is more likely to apply; there is a greater chance of the complete cases representing the population when only a few cases are missing. A researcher faces a more difficult decision when much data are missing as in the asthma study example. Approximately 88% of the informants fail to report one or more variables, leaving only 19 out of 154 cases for analysis. As seen earlier, some preliminary evidence from Little's (1988) MCAR test indicates that the assumption of MCAR data may not apply.

Even when the data are MCAR, complete case analysis has some potential difficulties. With large amounts of missing data (as may occur when several predictor variables are missing observations), relatively few cases may be left for the analysis as in the asthma data set. Complete case analysis may not be a viable option in all instances of MCAR data. The main advantage of the method is ease of implementation since the researcher can use standard methods for computing estimates for a proposed model. One disadvantage of the method centers on the number of cases that observe all variables of interest in the data; a researcher cannot anticipate if an adequate amount of data remain for the analysis.

Available Case Analysis

Available case analysis, or pairwise deletion, uses all available data to estimate parameters of the model. When a researcher looks at univariate descriptive statistics of a data set with missing observations, he or she is using available case analysis, examining the means and variances of the variables observed throughout the data set. This strategy is illustrated in Table 1 where we see descriptive statistics computed on different sets of cases.

When interest focuses on bivariate or multivariate relationships, the potential problems increase. Figure 1 illustrates a simple two-variable data matrix with only one variable subject to nonresponse. In pairwise deletion, all cases would be used to estimate the mean of X_1, but only the complete cases would contribute to an estimate of X_2, and the correlation between X_1 and X_2. Different sets of cases are used to estimate parameters of interest in the data. While Kim and Curry (1977) argue that estimates can be improved by using available cases instead of complete cases, others (Anderson, Basilevsky, & Hum, 1983; Haitovsky, 1968; Little, 1992; Little & Rubin, 1987) have pointed out problems with the procedure.

When the data are MCAR – when the remaining observations are representative of the originally identified data set – Little (1992) shows that available case analysis provides consistent estimates (the correct point estimates, see e.g., Cox & Hinckley, 1974) when variables are moderately correlated in regression models. When variables are highly correlated, available case analysis provides estimates that are inferior to complete case results as illustrated in simulations by Haitovsky (1968).

$$\left.\begin{array}{ll} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \bullet & \bullet \\ \bullet & \bullet \\ \bullet & \bullet \\ x_{1m} & x_{2m} \end{array}\right\} m \text{ Complete Cases}$$

$$\left.\begin{array}{ll} x_{1(m+1)} & - \\ \bullet & \bullet \\ \bullet & \bullet \\ \bullet & \bullet \\ x_{1n} & - \end{array}\right\}$$

n - m Cases with observations on x_1

Available Case Estimates:

$$\bar{x}_1 = \sum\nolimits_{i=1}^{n} x_{1i}$$

$$\bar{x}_2 = \sum\nolimits_{i=1}^{m} x_{2i}$$

$$s_1^2 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}{n-1}$$

$$s_2^2 = \frac{\sum_{i=1}^{m} (x_{2i} - \bar{x}_2)^2}{m-1}$$

$$r_{xy}^2 = \frac{1}{m-1} \frac{\sum_{i=1}^{m} (x_{1i} - \bar{x}_{1(m)})(x_{2i} - \bar{x}_2)}{s_{1(m)}\, s_2}$$

where $\bar{x}_{1(m)}$ and $s_{1(m)}$ are the mean and standard deviation of x_1 calculated from the m complete cases.

Fig. 1. Illustration of missing data restricted to one variable.

Another difficulty is that available case analysis can produce estimated covariance matrices that are implausible, such as estimating correlations outside of the range of -1.0 to 1.0. Errors in estimation occur because of the differing numbers of observations used to estimate components of the covariance matrix as illustrated in Figure 1 above. The relative performance of complete-case analysis and available case analysis, with MCAR data, depends on the correlation between the variables; available case analysis will provide consistent estimates only when variables are weakly correlated. The major difficulty with available case analysis lies in the fact that one cannot predict when available case analysis will provide adequate results, and is thus not useful as a general method.

Single-Value Imputation

In order to preserve the number of cases originally identified for the review, some researchers fill in the missing value with a plausible one, such as the mean for the cases that observe the variable. The analyst continues with the statistical method as if the data are completely observed. While this strategy allows the inclusion of all cases in a standard analysis procedure, replacing missing values with a single value changes the distribution of that variable by decreasing the variance that is likely present.

Little (1992) points out that while mean imputation results in overall means that are equal to the complete case values, the variance of these same variables is underestimated. This underestimation derives from two sources. First, filling in the missing values with the same mean value does not account for the variation that would likely be present if the variables were observed. The true values probably vary from the mean. Second, the smaller standard errors due to the increased sample size do not adequately reflect the uncertainty that does exist in the data. A researcher does not have the same amount of information present when some cases are missing important variables as he or she would have with completely observed data. Bias in the estimation of variances and standard errors are compounded when estimating multivariate parameters such as regression coefficients. Under no circumstances does mean imputation produce unbiased results.

Recommendations Concerning Widely Used Methods

Only complete case analysis provides valid estimates under the least number of conditions, and is thus applicable to a wider range of situations than available case analysis. Complete case analysis requires the assumption of

MCAR data, and enough complete cases to estimate the desired model. While available case analysis produces valid estimates with MCAR data, its properties are dependent on the magnitude of the correlations that exist between variables, thus limiting the method's applicability. Available case analysis also potentially produces invalid estimates due to the varying samples used to estimate parameters. There exist no systematic methods for determining when invalid estimates can occur with available case analysis, again limiting how generally the method can be applied. Mean imputation cannot be recommended under any circumstances.

Model-Based Methods for Multivariate Normal Missing Data

When a researcher decides that complete case or available case analysis will not provide adequate results, the researcher has alternative analysis procedures that come with the price of added assumptions about the distribution of the data and the nature of the missing data mechanism. The methods described in this section are called model-based methods since the researcher must make assumptions about the joint distribution of all variables in the model (including both outcomes and predictors). The discussion of the methods in this section is necessarily conceptual. Several works such as Little and Rubin (1987) and Schafer (1997) provide detailed and elegant treatments of the theory and analysis of missing data. The interested reader is referred to these references for the full exposition of the statistical theory underlying these methods. My goal is to introduce researchers to two model-based methods based on the multivariate normal distribution, and to provide suggestions for how a researcher might implement and interpret these methods. The major advantage of these two methods is that given the assumptions, the results obtained apply to a broader range of contexts with fewer conditions than the methods of the previous section. Though model-based methods require more complex computations, the availability of software in the near future should allow wider use.

The two model-based methods for missing data described in this section are maximum likelihood using the EM algorithm, and multiple imputation. The methods share two assumptions: that the joint distribution of the data is multivariate normal, and that the missing data mechanism is ignorable. Assuming that the data are multivariate normal is different from our typical assumptions about the relationship between an outcome variable and several predictor variables. For example, in a linear regression analysis, we assume that the outcome variable is univariate normal conditional on the values of the fixed predictors. For the model-based methods discussed here, we assume that the joint distribution of all variables, including outcomes and predictors is multivariate normal. Multivariate normal distributions have the advantage that any marginal, conditional distribution, such as that for predictor X_1, is a normal distribution conditional on all other variables in the model including the outcome variable, a property utilized in the estimation procedures. A disadvantage of the multivariate normal assumption concerns the inclusion of categorical variables. When nominal data are included in the model, we need to assume that these variables are completely observed, and that the remaining continuous variables are multivariate normal conditional on these categorical variables.

As discussed in a previous section, the assumption of an ignorable response mechanism requires MAR data. The assumption of an ignorable response mechanism is less restrictive than the assumption of MCAR data needed in complete-case analysis. The difficulty researchers have, however, lies in gathering empirical evidence about the response mechanism. The best we can do as researchers is to examine the sensitivity of results to different assumptions about the response mechanism. Comparing results from complete-case analysis and model-based methods may provide clues about the nature of the data. Even without direct evidence of MAR, the two model-based methods do have the advantage of using all information available in the data, a property that makes available case analysis seem reasonable. The next two sections provide a conceptual overview of two model-based methods for multivariate normal data.

Maximum Likelihood Methods Using the EM Algorithm

One method for estimating unknown parameters of a model is the use of maximum likelihood. When a researcher has complete data, quantities such as the mean and linear regression coefficients are easily computed, and are maximum likelihood estimates, based on maximizing the likelihood of the observed data. The same principle holds when missing data occur; we can base estimation on the likelihood of the observed data. The difficulty lies in specifying the likelihood of the observed data. The likelihood of the observed data is more complex when missing data occur than in our usual data analysis situation, and there is no simple solution to the problem of finding an estimate that maximizes the likelihood of the observed data. However, Dempster, Laird, and Rubin (1977) proposed the use of an iterative solution, termed the EM algorithm, to find the estimate of a parameter (such as the means and covariance matrix) when closed form solutions to the maximization of a likelihood are not possible. Both Little and Rubin (1987) and Schafer (1997) provide the theory of the EM algorithm for missing data analysis assuming multivariate normal data. Here, I will provide a conceptual overview of the concepts.

When missing data occur, the observed data likelihood is difficult to maximize. For example, when we have completely observed data, the maximum likelihood estimate of the mean is the sum of the values of a particular variable divided by the sample size. When some values of that variable are missing, we do not know the value of the sum, and thus cannot compute the value of the estimate of the mean. The EM algorithm splits up the problem into two estimation problems. For the first estimation, imagine that we knew the value of the population mean and variance. With the value of the population mean and variance, we have the specification of the population distribution. From the population distribution, we could estimate the expected value of the sum of a variable, which is the mean of the variable multiplied by the number in the sample. For the second estimation problem, if we knew the value of the sum of a variable, we could get an estimate of the population mean (and with the sums of squares and crossproducts, we could get the covariance matrix). The estimation or E-step of the EM algorithm computes the expected value of the sum of the variables with missing data assuming that we have a value for the population mean, and variance-covariance matrix. The maximization, or M-step, uses the expected value of the sum of a variable to estimate the population mean and covariance. The process cycles back and forth until the estimates do not change substantially. In each step, we assume that we know one of the two unknown pieces of information (e.g., either the value of the population mean we are estimating, or the sum of the variable), and then compute the other parameter.

Maximum likelihood methods for missing multivariate normal data focus on the estimation of the parameters of the observed data, namely the mean vector and variance-covariance matrix. Because we assume the data multivariate normal, we can utilize the well-known properties of conditional normal distributions to estimate the expected values of the sums and crossproducts of the variables. Using maximum likelihood with the EM algorithm does not result in values for individual missing variables. The researcher obtains estimates for the means and the variance-covariance matrix of the variables of interest, and then uses these parameter estimates to obtain model parameters such as the coefficients of a linear regression model.

As the reader may guess, estimation of parameters using the EM algorithm requires some specialized computing. Both BMDP (Dixon, 1992) and SPSS (1999) have modules that compute the maximum likelihood estimates of the mean vector and covariance matrix of a multivariate normal data set. The program NORM written by Schafer (1999) also computes the maximum likelihood estimates. The one major difficulty with all of the above-mentioned programs is the computation of the standard errors of estimates (such as the standard error of the mean). Testing whether a mean is significantly different from zero, for example, requires an estimate of how accurate our estimate is. In maximum likelihood theory, the negative second derivative of the observed data loglikelihood is needed to obtain standard errors of the estimated mean vector and covariance matrix. This quantity requires algebraic analysis to compute, and is unique to every set of multivariate data. Another method for obtaining the standard errors of parameters when missing data occur has been suggested by Meng and Rubin (1991). Meng and Rubin's method entails adding steps to the EM algorithm code, and requires specialized computing knowledge. Unfortunately, the standard errors given in BMDP and SPSS are not those computed using the negative second derivative nor the process suggested by Meng and Rubin. Without adding to the EM code, the best a researcher can do is to examine maximum likelihood estimates with a conservative estimate of standard errors such as those obtained from complete case analysis. The computation difficulties and inability to compute standard errors does limit the usefulness of this method.

Multiple Imputation with Multivariate Normal Data

Multiple imputation avoids two of the difficulties associated with maximum likelihood methods using the EM algorithm. With multiple imputation, a researcher will use standard methods of analysis once imputations are computed, and can easily obtain standard errors of estimates. Though specialized computing is required in multiple imputation, the method provides much more flexibility than in the method described in the previous section. Rubin (1987, 1996) and Schafer (1997) provide comprehensive discussions of the statistical theory of multiple imputation, as well as several examples of its use. The discussion here will be conceptual, and the reader is urged to examine these texts for more detailed information.

In multiple imputation, the researcher generates several possible values for each missing observation in the data in order to obtain a set of three or more parallel completed data sets. Using standard analysis procedures, the researcher analyzes each completed data set, and then combines these estimates to obtain the multiply-imputed estimates. Thus, multiple imputation has two stages: the generation of the parallel completed data sets, and the computation of the multiply-imputed estimates. The major work involved in multiple imputation involves generating possible values for each missing observation.

The set of possible values for missing observations are based on the distribution of the data, in this case the multivariate normal. Unlike maximum likelihood, the goal of multiple imputation is to obtain estimates of the missing values rather than the expected values of the sufficient statistics. Estimates of missing values are obtained by simulating random draws from the distribution of the missing variables given the observed variables. Distributions of the data are derived from Bayesian theory, so that the researcher samples values from the posterior probability distribution of the missing values given the observed variables. Assuming a multivariate distribution, a distribution whose properties are well-known, provides a way to posit the distribution of a variable missing observations, and then to draw random samples from that distribution as possible values for the missing value. As in maximum likelihood, the posterior probability distribution of the missing variables given the observed variables is complex, and requires a two-step algorithm, referred to as data augmentation, with a logic similar to that of the EM algorithm.

As in maximum likelihood with the EM algorithm, we are interested in estimating the parameter of the joint distribution of the data, such as the mean vector and covariance matrix. If we knew the values of the missing observations, we could easily estimate θ using well-known estimates (such as the sum of a variable divided by the sample size for the mean). Similarly, if we knew all the means and covariance matrix, then we could specify the marginal distribution of the missing values given the observed data, and sample plausible values for the missing data from that distribution (akin to drawing names out of a hat). The first step of the estimation process is to assume θ is known, and to sample values for the missing observations from the conditional distribution of the missing data given the observed data and the current estimate of θ. The second step is to assume that we have complete data, and re-estimate θ. The estimation procedures cycle back and forth between these two steps.

The discussion above represents a simplistic view of the estimation process. Schafer (1997) and Tanner (1993) discuss methods for data augmentation in detail. Data augmentation requires several hundred and sometimes thousands of iterations before one can begin to save completed data sets for analysis.

Each iteration of the data augmentation algorithm as implemented in the example of this paper generates one completed data set, so that the researcher may go through thousands of iterations to obtain three to five completed data sets for analysis. The number of iterations needed in a given application depends on the fractions of missing data in any given analysis. Schafer (1997) provides a number of convergence diagnostics that help a researcher assess whether the number of iterations is sufficient; these diagnostics will be illustrated in the example.

A second issue centers on the number of completed data sets needed for multiple imputation. Schafer (1997) discusses why $m = 5$ completed data sets typically result in unbiased estimates. Once the researcher has the *m* completed data sets, he or she performs the targeted statistical analysis on each data set separately, obtaining *m* estimates of the parameters of the model. These parameters may be regression coefficients of a linear model, the means and covariance matrix of the data, or any other set of model parameters. These estimates are combined using rules discussed by both Rubin (1987) and Schafer (1997). Schafer (1999) provides a stand-alone program, NORM, that will generate the multiple imputations for a multivariate normal data set. The researcher inputs these data sets into another statistical computing program of the researcher's choice, obtaining *m* estimates of the target parameters. The researcher then inputs the *m* estimates of the model parameters into the NORM program, and can obtain the multiply-imputed estimates along with the standard errors of these estimates.

The assumptions of multivariate normality and an ignorable response mechanism apply to multiple imputation as well as maximum likelihood using the EM algorithm. The same difficulties apply as well; the multivariate normal assumption makes the use of non-ordered categorical variables problematic, and evidence about the ignorability of the response mechanism is difficult to obtain. Again, comparing results from several analyses may provide clues about the nature of the data. Multiple imputation provides an added feature. A researcher could generate imputations using alternative assumptions about the response mechanism generating imputations using a censoring mechanism, for example. Though this strategy may require specialized computing, the researcher could systematically examine the sensitivity of results to assumptions about the nature of the missing data mechanism.

How do the results of multiple imputation and maximum likelihood with the EM algorithm compare? Both multiple imputation and maximum likelihood are based on the likelihood theory. Multiple imputation results do converge to the maximum likelihood results given the data model used for both is the same. Computationally, multiple imputation is easier to implement with the introduction of specialized computer packages as well as planned modules in both SPSS and SAS (Barnard, 2000). Analytically, multiple imputation provides more flexibility. Maximum likelihood with the EM algorithm results in point estimates of the mean vector and variance-covariance matrix. Any subsequent analysis of the data, such as estimating parameters of a linear regression model, are computed as functions of these means and variances. Multiple imputation results in completed data sets so that the analysis consists of performing standard analyses on more than one data set, and then combining those estimates into multiply-imputed estimates. From these multiple imputation data sets, one can conduct different analyses. Thus, multiple imputation is particularly useful in large public use data sets; once imputations are performed, different analysts can use the data to explore different models. As Barnard (2000) points out, multiple imputation separates the task of imputing the data from analyzing the completed data sets so that researchers using the data do not have to handle difficulties from the missing data as well as analyze the data.

While multiple imputation appears the most promising of current missing data methods, Rubin (1996) critically reflects on the use of multiple imputation over the past 20 years. Some criticisms of the method center on the amount of computing and analysis time. Analyzing five sets of data is certainly more costly than one analysis, and the method does require specialized software. A more critical assessment comes from Fay (1991, 1992, 1993, 1994, 1996) and is also addressed by Meng (1994). Fay focuses on the use of multiple imputation in large, public-use data sets where the person imputing the data is separate from the analyst. Whatever model the analyst fits using the imputed data sets must be congenial to (must include the same variables) as the model used by the person who originally imputed the data. As Rubin (1996) notes, the model used to generate the multiple imputations should include all the variables likely to be used in subsequent analyses. The statistical literature on missing data continues to grow, and we in the social sciences are bound to have more tools at our disposal in the future.

Data Analysis Examples using Complete Cases, Maximum Likelihood and Multiple Imputation

For the example, I use the asthma study data as described in Tables 1 and 2. The analysis examines the relationship of a student's self-efficacy in controlling their asthma (Belief) and the following predictors: participation in the treatment or control group (Group), self-rating of symptom severity for a 2-week period post-treatment (Symsev), reading test score on a state standardized test (Reading), age of child in years (Age), gender of child (Gender), and number of allergies reported by child's parent (Allergy).

Complete and Available Case Analyses

As illustrated in Table 2, there are 19 cases with complete data on all the variables of interest. When we use a statistical package like SPSS, the default procedure when missing data occurs is listwise deletion, so that only the 19 cases are used for the analysis. The complete case results are given in the first two columns of Table 3. The regression coefficients for the intercept and Reading are significantly different from zero at $p < 0.05$; Group, Symsev and Age have regression coefficients that are significantly different from zero at $p < 0.10$. Each Reading grade equivalent score increase is associated with almost one-half of a point increase on the Belief scale. The students in the control group tend to score one-half of a point lower than treatment students on the Belief scale, indicating a moderate effect of the treatment in raising self-efficacy levels while controlling for all other variables. Those with more severe symptoms also score lower. Students whose average symptom ratings were higher (indicating more severe symptoms) over the course of the 2 weeks after treatment tended to have a lower score on self-efficacy. Age has a puzzling relationship with Belief score; older children tend to score lower on the Belief scale than younger children. The correlation between Reading score and Age is 0.658, a fairly large correlation. The negative coefficient could relate to collinearity between Reading score and Age, though the two variables are not perfectly correlated. From Table 1, we see that the grade equivalent Reading scores for the students in the sample span a wider range (1st through 8th grades) while the age range would imply students in the third through ninth grades. Many students are not reading at their grade level.

The available case analysis, presented here for comparison, provides a different set of model estimates. None of the predictors are significantly different from zero. The R^2 value is given as 0.215, with an adjusted R^2 of -0.147, an implausible value. The available case results would seemingly not fit the data well. Inferences about the model from this analysis would also lead to differences from the complete case model. For example, the treatment group difference is no longer close to statistical significance, and Reading

Table 3. Results.

	Complete Case Analysis (N = 19)		Available Case Analysis (N = 19)		Maximum Likelihood Analysis (N = 154)		Multiple Imputation Analysis (N = 154)		
Variable	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*	*% Mis. Info.*
Intercept	4.617*	0.838	3.794*	1.219	4.083	0.362	3.994*	0.412	24.0
Trt group	−0.550	0.276	−0.038	0.388	−0.132	0.112	−0.119	0.123	17.7
Symsev	−0.315	0.161	−0.485	0.503	−0.480	0.144	−0.498*	0.147	1.6
Reading	0.409*	0.096	0.171	0.132	0.218	0.039	0.201*	0.046	40.7
Age	−0.211	0.115	−0.025	0.139	−0.089	0.043	−0.067	0.053	41.8
Gender	0.198	0.189	0.053	0.364	0.084	0.104	0.046	0.110	10.6
Allergy	−0.005	0.057	0.018	0.099	0.063	0.029	0.045	0.053	75.6

level is not associated with Belief score. All of the standard errors for the regression coefficients are larger than in the complete case analysis.

Checking the Multivariate Normality Assumption

For the two model-based procedures, we need to assume that the data are multivariate normal, and that the response mechanism is ignorable. To examine the normality assumption, Figure 2 shows the histograms of each of the continuous variables. Three of the predictors – Age, Reading and Allergy – all have fairly normal distributions. The histograms for both Symsev and Belief are skewed. Both variables suffer from range restrictions. Many of the students did not experience symptoms in the 2 weeks post-treatment, while a few did have active asthma episodes. The opposite was true for the belief scale. On average, most students rated their self-efficacy for handling their asthma as high; the average item rating was close to the maximum value of 4, meaning that students felt much confidence in handling their symptoms. The two categorical data variables in the data, Group and Gender are completely observed. Within the four cells defined by these two variables, we need to assume that the variables have normal distributions, or in other words, that variables in the model must have normal distributions conditional on these two variables. The histograms by Group and Gender are similar to those presented in Figure 2.

Results Using Maximum Likelihood with the EM Algorithm

I used the program NORM (Schafer, 1999), to obtain the maximum likelihood estimates of the mean and correlation matrix. Alternatively, one could also use BMDPAM (Dixon, 1992) or SPSS Missing Values Analysis (SPSS, 1999) to compute the maximum likelihood estimates of the mean and covariance matrix. After obtaining the maximum likelihood estimates, I used SPSS (1999) to estimate the parameters of the linear regression model from the mean vector and correlation matrix. The SPSS syntax for computing the regression estimates from the mean vector and correlation matrix is given in the Appendix.

The middle right two columns of Table 3 provide the maximum likelihood estimates for the linear model and the standard errors obtained from SPSS. Note that the standard errors for the estimates will be biased, since they are not based on the negative second derivative of the loglikelihood, nor do they use Meng and Rubin's (1991) SEM algorithm. The standard errors are computed as if the mean and correlation matrix imported into the regression module are

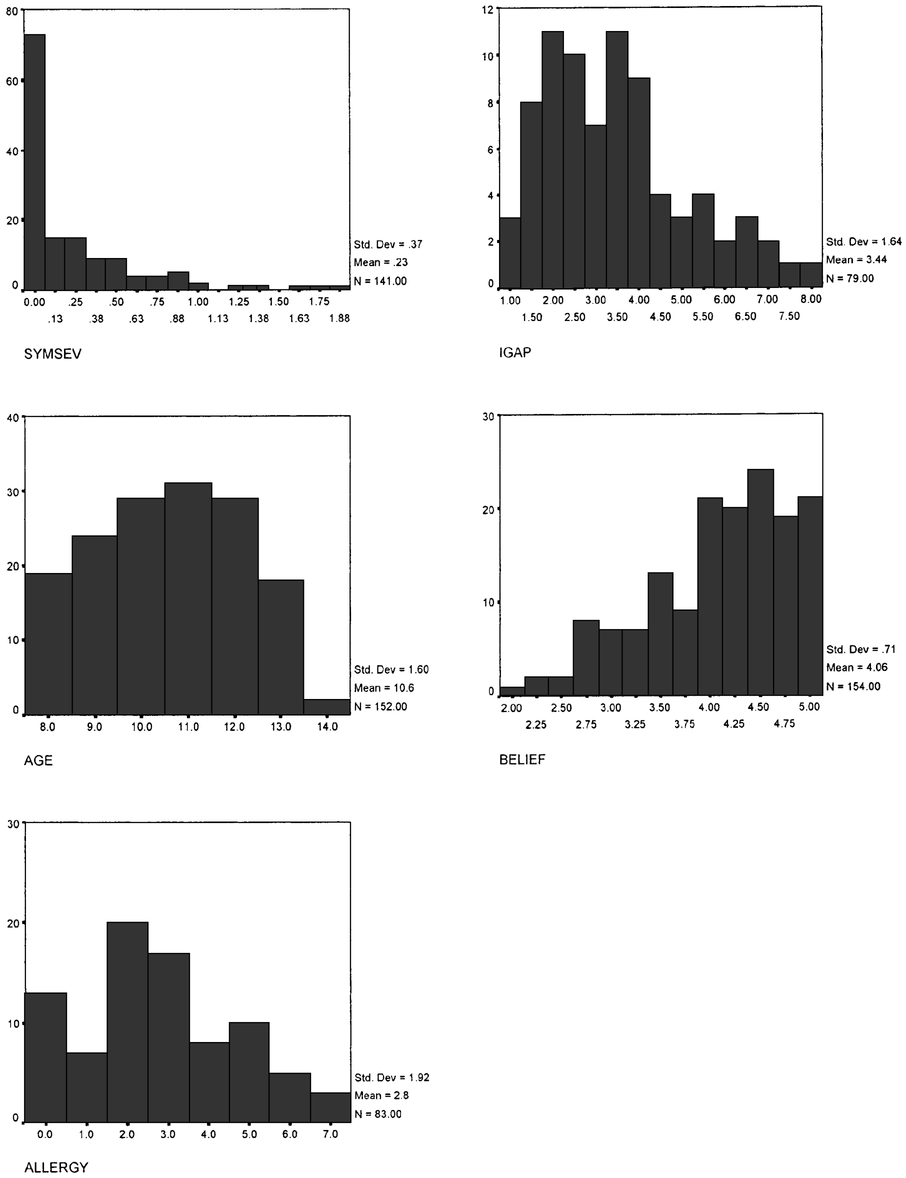

Fig. 2. Histograms of variables

from a complete data set. Looking at Table 3, we see that the standard errors computed by the standard regression program are smaller than from the complete case analysis given that the errors are computed with 154 cases. We could utilize the more conservative standard errors from complete cases to give a rough approximation of which estimates are significantly different from zero. This standard would suggest that the intercept, Symsev, and Reading may have associations with Belief that are worth interpretation. The maximum likelihood estimates suggest a somewhat stronger relationship between ratings of symptom severity and self-efficacy; every increase in average ratings of severity of symptoms results in an almost one-half of a point decrease in self-efficacy beliefs. Reading score is moderately related to self-efficacy beliefs though the magnitude of the relationship is about half of that estimated by the 19 complete cases. Using maximum likelihood, we do not find a significant difference between the treatment and control groups though the direction of the difference favors the treatment. Age is not associated with Belief score in the maximum likelihood analysis.

Multiple Imputation Results

For multiple imputation, I used Schafer's (1999) NORM program. At present, only Schafer's program provides multiple imputation that is based on the statistical theory discussed in Schafer (1997). As Schafer (1997) and Tanner (1993) discuss, the data augmentation procedures used to compute the multiple imputations can produce sample estimates that are correlated with each other in early rounds of iterations, before estimates converge. Researchers using data augmentation must check the correlations between adjacent iterations to ensure independent samples are obtained, and must allow the algorithm to run for several hundred iterations. After 20,000 iterations of data augmentation using a noninformative prior distribution, estimates from adjacent iterations for Allergy were still highly correlated indicating that the data augmentation algorithm had not converged. Schafer (1997) discusses the uses and interpretations of the diagnostic plots in detail. Looking at Table 2, we see that almost half of the cases are missing an observation for Allergy. As suggested by Schafer (1997), I then standardized all the variables in the data set to have a standard normal distribution, and then used a ridge prior, what in Bayesian terms is called a weak prior. Using this prior distribution can stabilize the estimation procedures. After 20,000 iterations, the adjacent estimates of all the variables including Allergy showed little or no correlation. I then ran a second chain of the data augmentation algorithm with the ridge prior, saving the augmented data sets at iterations 22,000; 24,000; 26,000; 28,000 and 30,000. Using Schafer's NORM program, the iterations took about 2 minutes to complete.

To obtain the multiply-imputed estimates of the linear regression model, I entered each completed data set into SPSS (1999) to obtain five sets of estimates. Schafer's NORM program (1999) provides the multiply-imputed estimates for the linear regression model using the combining rules discussed by both Schafer (1997) and Rubin (1987). The last three columns of Table 3 provide information on the multiple imputation analysis. Note that all of the standard errors of the estimates of the regression coefficients have values that lie between those based on the 19 complete cases, and those based on the incorrect assumption in the maximum likelihood analysis that all 154 cases are complete. The intercept, Symsev, and Reading have estimates that are significantly different from zero. As in the maximum likelihood analysis, those students with higher ratings for the severity of their asthma symptoms tended to score about one-half a point lower on self-efficacy, an estimate larger than the complete case analysis. Higher reading scores are also associated with a small increase in Belief scores though as in the maximum likelihood case the estimate is half as large as in complete cases.

Neither treatment group nor age is associated with Belief score in either multiple imputation or in maximum likelihood. The treatment group effect (Group) is not significantly different from zero in either maximum likelihood or multiple imputation. The two model-based methods also do not observe any significant relationship with Age.

The last column of Table 3 provides another diagnostic important for interpreting the results of the maximum likelihood and multiple imputation analyses. Both Rubin (1987) and Schafer (1997) discuss the derivation and justification of the percent of missing information for each variable. Conceptually, the percent or fraction of missing information gives an estimate of how much information we do not have about the parameters of interest given the amount of missing data. This diagnostic is based on the ratio of the variance we estimate for a given parameter from multiple imputation and the estimated variance if we had no missing data on that parameter. Three variables have percentages of missing information near 50%, indicating that the data set provides a small amount of information about how Reading, Age, and Allergy relate to the other variables in the model. The percentage of missing information on these variables reiterates the fact that both Reading and Allergy have much missing data. The correlation between Age and

Reading and the narrow range of Age in the sample may contribute to the large amount of missing data we estimate for Age even though we have almost complete data on that variable.

DATA INTERPRETATION

The small number of cases with complete data pose a problem in this intervention study. Given that 154 students participated in the study, using only 19 cases to represent the whole does not appear justified; we do not believe that the 19 cases are a random sample from the whole data set though we do not have direct evidence of this assumption. Instead, we can make a somewhat less demanding assumption by considering the data MAR. With MAR data, we can use maximum likelihood methods or multiple imputation. Assuming the data multivariate normal, we obtain estimates for our regression model. As anticipated, there are few differences in the magnitude of the estimates from maximum likelihood and multiple imputation since multiple imputation methods converge to those from maximum likelihood. The difference, however, is in the computation of the standard errors as well as in overall ease of computation. Programs such as NORM (Schafer, 1999) provide both maximum likelihood and multiple imputation results, and more programs are likely to be available in the near future (Barnard, 2000).

The interpretation of the models do differ between the model-based and complete case estimates. The treatment group did not score appreciably higher on the measure of self-efficacy while controlling for reading ability, gender, age, severity of symptoms, and overall number of allergies suffered. Children who do report more severe symptoms, however, do tend to doubt their ability to control their asthma. The study does, however, suffer from a common problem – much incomplete data on important measures. The number of allergies suffered by a child provides one measure of the risks that the child may suffer an acute episode, especially if the asthma is not under control. Children in this population attended schools in the inner-city, and are likely to frequent environments where their allergies and asthma symptoms are exacerbated. An alternative measure to Allergy of a student's risk for asthma episodes may provide more information for examining the model. In addition, the inclusion of both Age and Reading may result in collinearity problems in the model itself. However, using both in the model to create the multiple imputations is warranted given the correlation between these two variables.

The Reading variable may also be one more likely to be missing for reasons that are nonignorable; schools with low reading scores may have been less likely to share that information with the researchers. One way to increase the possibility of MAR data with the Reading score is to include more variables in the imputation stage that could relate to the ability of a student to read and interpret the Belief scale such as Age or grades in school. The data set does contain many other health outcomes and attitude data that may prove useful in improving our estimates of the relationships between Allergy and Belief and between reading ability and Belief score.

While the available case analysis provides point estimates close to those from model-based methods, we cannot be sure under what conditions available case analysis will provide unbiased estimates. Both Schafer (1997) and Barnard (2000) refer to model-based methods as principled methods, meaning that they are based on statistical theory, and can be shown to produce unbiased estimates when the assumptions of the method are met. Available case analysis is not a principled method, and thus cannot be relied on to produce unbiased estimates in a wide range of missing data problems.

RECOMMENDATIONS

A significant amount of research has appeared in the statistics literature to convince social scientists to use methods other than available case and mean imputation to handle missing data. When few cases are missing values, complete case analysis methods can provide unbiased estimates. In other circumstances, as in the asthma intervention study, the number of complete cases is a small fraction of the total. The expense and investment in the study warrant our using methods that utilize as much data as possible.

Social scientists do face major hurtles in implementing innovations that occur in the statistical literature. We often do not have the needed expertise or technology for implementing these methods. Even when software is available, we often witness the misuse of methods because of the ease of computations. Complete case analysis as well as available cases and mean substitution are used most frequently since they are included as options in many statistical packages. However, as others have warned repeatedly (Little & Rubin, 1987; Rubin, 1987; Schafer, 1997), these methods do not provide adequate results in all instances, and in fact, available case and mean substitution provide problematic estimates in almost all instances. When missing data occur, we need to acknowledge the limitations of our data. Model-based methods provide this acknowledgment through added assumptions about the nature of the missing data, and the greater effort required to analyze the data.

One of the major contributions of the missing data literature is the emphasis on explicitly stating the assumptions used to analyze the data. While we cannot necessarily provide empirical evidence for the appropriateness of the assumption, we can openly discuss the assumption with the final judgement left to the readers. We can also examine the sensitivity of our results to differing assumptions. In our asthma intervention example, the complete case analysis differs from those using model-based procedures. Given that my assumptions for the model-based procedures (namely MAR data and multivariate normality) are less restrictive than the complete case analysis assumptions (MCAR data), I am inclined to accept the results from the model-based procedures.

The likely introduction of missing data methods in major statistical packages should allow the greater use of the methods discussed here, as well as for methods dealing with more complex forms of missing data. The use of model-based methods will aid in the analysis of much social science data, data which are often incomplete and missing. I also hope that the availability of software for missing data analysis brings an increased willingness by researchers to state explicitly the assumptions used in the analysis and to explore the sensitivity of results to those assumptions. Both of these innovations will serve to strengthen our research efforts.

ACKNOWLEDGEMENT

I wish to thank Barbara Velsor-Friedrich for the use of a subset of her data. I also thank Rebecca Barr, David Prasse, and anonymous reviewers for helpful comments on earlier drafts of the paper.

REFERENCES

Anderson, A.B., Basilevsky, A. & Hum, D.P. (1983). Missing data: A review of the literature. In J.D. W.P.H. Rossi & A.B. Anderson (Eds.), *Handbook of survey research*. New York: Academic Press.

Barnard, J. (2000, June). *Multiple imputation for missing Data.* Paper presented at the Summer Workshop of the Northeastern Illinois Chapter of the American Statistical Association, Northbrook, IL.

Cohen, J., & Cohen, P. (1975). *Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences*. New York: John Wiley.

Cox, D.R., & Hinckley, D.V. (1974). *Theoretical Statistics*. London: Chapman & Hall.

Dempster, A.P., Laird, N.M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Association, B39*, 1–38.

Dixon, W.J. (Ed.). (1992). *BMDP Statistical Software Manual* (Vol. 2). Berkeley, CA: University of California Press.

Fay, R.E. (1991). *A Design-Based Perspective on Missing Data Variance.* Paper presented at the 1991 Annual Research Conference, U.S. Bureau of the Census.

Fay, R.E. (1992). *When are Inferences from Multiple Imputation Valid?* Paper presented at the Section on Survey Research Methods, American Statistical Association.

Fay, R.E. (1993). *Valid Inferences from Imputed Survey Data.* Paper presented at the Section on Survey Research Methods, American Statistical Association.

Fay, R.E. (1994). *Analyzing Imputed Survey Data Sets With Model-Assisted Estimators.* Paper presented at the Section on Survey Research Methods, American Statistical Association.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association, 91*, 490–498.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society, B30*, 67–82.

Heitjan, D.F., & Basu, S. (1996). Distinguishing 'Missing at Random' and 'Missing Completely at Random'. *American Statistician, 50*, 207–213.

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.

Kim, J., & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods & Research, 6*, 215–240.

Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202.

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., & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Sciences, 9*, 538–573.

Meng, X., & Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86*, 899–909.

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

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

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473–489.

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

Schafer, J.L. (1999). *NORM*: *Multiple imputation of incomplete multivariate data under a normal model, version 2*. Software for Windows 95/98/NT, available from http://www.stat.psu.edu/~jls/misoftwa.html.

SPSS. (1999). SPSS for windows (Version Rel. 9.0). Chicago: SPSS, Inc.

Tanner, M.A. (1993). *Tools for Statistical Inference*. New York: Springer-Verlag.

Velsor-Friedrich, B. (in preparation). Results of an asthma intervention program in eight inner-city schools.

APPENDIX

SPSS Syntax for Linear Regression Using Means and Correlation Matrix

```
* Matrix data with procedure REGRESSION.
MATRIX DATA VARIABLES=group symsev igap age gender belief2 allergy
/CONTENTS=MEAN SD N CORR /FORMAT=LOWER DIAGONAL.

BEGIN DATA
0.558  0.244  4.062  10.578  .442  4.057  2.869
0.496  0.370  1.790  1.597  0.496  0.711  1.908
154  154 154 154 154 154 154
1.0
-0.318   1.0
0.276  -0.173  1.0
0.175  -0.128  0.611  1.0
-0.157  -0.0271  0.0621  0.121  1.0
0.0752  -0.296  0.418  0.188  0.0801  1.0
-0.116  -0.0262  -0.184  0.170  -0.0566  0.0476  1.0
END DATA.

REGRESSION  Matrix = in(*)
/VARIABLES=group  symsev igap age gender
belief2  allergy
/criterion tolerance(.0000000001)
/DEP=belief2  / method=ENTER.
```

The / CONTENTS line provides the format of the data. The first line after the BEGIN DATA line contains the maximum likelihood estimates of the means, the second line, the maximum likelihood estimates of the standard deviations of the variables, the sample size used for each estimate (here the hypothetical complete sample size), and the lower half of the maximum likelihood estimate of the correlation matrix.