Excerpts from Missing Data, by Paul Allison.
Paper # 136 in the Sage Series on
Quantitative Applications in the Social
Sciences. 2002. Sage Publications:
Thousand Oaks.

MISSING DATA

PAUL D. ALLISON
University of Pennsylvania

1. INTRODUCTION

Sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data. In a typical data set, information is missing for some variables for some cases. In surveys that ask people to report their income, for example, a sizable fraction of the respondents typically refuse to answer. Outright refusals are only one cause of missing data. In self-administered surveys, people often overlook or forget to answer some of the questions. Even trained interviewers occasionally may neglect to ask some questions. Sometimes respondents say that they just do not know the answer or do not have the information available to them. Sometimes the question is inapplicable to some respondents, such as asking unmarried people to rate the quality of their marriage. In longitudinal studies, people who are interviewed in one wave may die or move away before the next wave. When data are collated from multiple administrative records, some records may have become inadvertently lost.

For all these reasons and many others, missing data are a ubiquitous problem in both the social and health sciences. Why is it a problem? Because nearly all standard statistical methods presume that every case has information on all the variables to be included in the analysis. Indeed, the vast majority of statistical textbooks have nothing whatsoever to say about missing data or how to deal with it.

There is one simple solution that everyone knows and that is usually the default for statistical packages: If a case has any missing data for any of the variables in the analysis, then simply exclude that case from the analysis. The result is a data set that has no missing data and can be analyzed by any conventional method. This strategy is commonly known in the social sciences as *listwise deletion* or *casewise deletion*, but also goes by the name of *complete case analysis*.

In addition to its simplicity, listwise deletion has some attractive statistical properties, which will be discussed later. It also has a major

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.