The Treatment of Missing Data

David C. Howell
University of Vermont

The treatment of missing data has been an issue in statistics for some time, but it has come to the fore in recent years. The current interest in missing data stems mostly from the problems caused in surveys and census data, but the topic is actually much broader than that. In this chapter I will discuss the treatment of missing data across a range of experimental designs, starting with those designs whose treatment is relatively straightforward (though not necessarily satisfactory) and moving to situations where the optimal solution is elusive. Fortunately we have come a long way since someone, I forget who, could say that the best treatment for missing data is not to have any. That may be the best way, but recent techniques have come far in narrowing the gap between the ideal and the practical.

The treatment of missing data is not an area that is particularly controversial. There are a number of alternative approaches, but there is pretty much universal agreement about the strengths and weaknesses of each. Over time new procedures replace older ones, but this, like many areas in statistical methods, is an area that changes slowly. So we often find that the older methods are still used, but that is mostly because in a specialized area like this, it takes a long time for newer methods to be understood and to push out the old.

My goal in this chapter is to give the reader an understanding of the issues involved in the treatment of missing data and the ability to be conversant with the approach that is adopted. When it comes to selecting an approach, it is not necessary to have an in depth knowledge of the technical issues, but it is necessary to understand the alternatives and to have a grasp of what is involved in each method.

Type of Missingness

Any discussion of missing data must begin with the question of why data are missing in the first place. They could be missing for perfectly simple and harmless reasons, such as a participant having an automobile accident and not being able to appear for testing. In such a case missingness is more of a nuisance than a problem to be overcome. On the other hand, data could be missing on the basis of either the participant's potential score on the dependent variable (Y) or any of the independent variables (X_i). The reasons for missing data plays an important role in how those data will be treated.

Missing Completely at Random (MCAR)

Rubin (1976) defined a clear taxonomy of missingness that has become the standard for any discussion of this topic. This taxonomy depends on the reasons why data are missing. If the fact that data are missing does not depend upon any values, or potential values, for any of the variables, then data are said to be **missing completely at random (MCAR).** The example above of the careless motorist who does not appear for testing because of an accident having nothing to do with the study is a case in point. Pickles (2005) phrased the condition somewhat differently by saying that for MCAR the probability of missingness is a constant. Any observation on a variable is as likely to be missing as any other. If you are going to have missing data, this is the ideal case because treatment of the existing data does not lead to bias in the estimated parameters. It may lead to a loss in power, which is often not a serious problem in census work, though it certain can be in experimental studies, but it won't lead to biased parameter estimates.

Little (1998) has provided a statistical test of the MCAR assumption. His MCAR test is a chi-square test. A significant value indicates that the data are not MCAR. This test is provided in the SPSS Missing Values Analysis (MVA), and should be applied whenever there is some question about MCAR. SAS also includes this test in PROC MI.

Missing at Random (MAR)

Data are **missing at random** if the probability of missing data on a variable, (Y), is *not* a function of its own value after controlling for other variables in the design. Allison (2001) uses the example of "missingness" for data on income being dependent on marital status. Perhaps unmarried couples are less likely to report their income than married ones. Unmarried couples probably have lower incomes than married ones, and it would at first appear that missingness on income is related to the value of income itself. But the data would still be MAR if the conditional probability of missingness were unrelated to the value of income *within each marital category*. Here the real question is whether the value of the variable determines the probability that it will be reported, or whether there is some other variable (X) where the probability of missingness on Y is conditional on the levels of X. To put it more formally, data are MAR if $p(Y \text{ missing } | Y,X) = p(Y \text{ missing } | X)$.

Missing Not at Random

Data are classed as **missing not at random (MNAR)** if either of the above two classifications are not met. Thus if the data are not at least MAR, then they are missing not at random. When data are MNAR there is presumably some model that lies behind missingness. If we knew that model we might be able to derive appropriate estimators of the parameters in the model underlying our data. Unfortunately we rarely know what the missingness model is, and so it is difficult to know how to proceed. In addition,

incorporating a model of missingness is often a very difficult task and may be specialized for each application.

Ignorable and Nonignorable Missingness

As I have suggested, when we have data that are MNAR, life becomes very much more difficult. Here we say that the mechanism controlling missing data in **nonignorable**. That means that we cannot sensibly solve whatever model we have unless we also are able to write a model that governs missingness. Modeling missingness is a very difficult thing to do, and most discussions, including this one, do not discuss the treatment of data whose missingness is nonignorable.

On the other hand, if data are at least MAR, the mechanism for missingness is **ignorable**. Thus we can proceed without worrying about the model for missingness. This is not to say that we can just ignore the problem of missing data. We still want to find better estimators of the parameters in our model, but we don't have to write a model that gets at missingness. We certainly have enough to do to improve estimation without also worrying about why the data are missing.

Missing Data and Alternative Experimental Designs

How we deal with missing data depends in large part on the experimental design that we are employing. Consider the difference between a correlational study where data on many variables are collected and then subjected to an analysis of linear regression, and an experimental study where we have two independent variables, usually categorical in nature, and one dependent variable. In an analysis of variance setting we most often think in terms of "unequal sample sizes" rather than "missing data," although unequal sample sizes very often are the direct result of data being missing rather than a planned inequality. With unequal sample sizes the techniques are quite well worked out. But in regression, we often want to substitute pseudovalues of a variable (referred to hereafter as "imputing data") and then solve the regression with a complete data set. The way we approach these two examples is quite different.

Traditional Experimental designs

For those whose main focus is experimental studies of behavior, the idea of missing data usually means that a person did not show up to participate in a study, or that one classroom had more students than another, or that a piece of equipment didn't record the data correctly. In these situations missing data create problems, but they are nothing like the problems involved in survey research, for example. In this section I am not taking a strict interpretation of "experimental" by always requiring that observations be assigned

to levels of the independent variable(s) at random. But I am distinguishing those studies which we loosely call "experimental" from those that we think of as "observational."

In experimental studies we most often have data missing on the dependent variable, though there are times when it is the independent variable that is missing. The latter situation is most common when the level of the independent variable is defined by self-report on the part of the participant, though there can be other causes. We have a somewhat different problems depending on whether it is the independent or dependent variable that is missing

Missing data on the independent variable

We will begin with the situation in which we class observations into groups on the basis of self-report, and then try to compare those groups on some dependent variable. For example, Sethi and Seligman (1993) compared three religious groupings, "Liberal, Moderate, and Fundamentalist," on their level of optimism. In their study they were able to identify religious groups on the basis of direct observation though obviously random assignment was not an option. But what if they had identified groups on the basis of a separate item that they included on the questionnaire that they gave out to measure optimism? Certainly there are a number of people who would fail to disclose their religious affiliation, and it is unlikely that the probability of disclosure is constant across all religious groups. (Certainly if you consider your religious affiliation to be a local coven, you would probably be less likely to report it than if you went to the nearby Methodist church.)

In this situation the simplest approach is to form four groups instead of three. In other words we identify participants as Liberal, Moderate, Fundamentalist, and Missing, and then run the analysis using those four groups. If contrasts show that the Missing group is not different from the other groups, we might be justified in dropping the missing data and proceeding normally with our analysis of the other three groups.

If we discover that the mean optimism score from those participants for whom group membership is unknown is significantly different from some other means (but perhaps not from all), we have a problem of interpretation, but at least we have learned something about missingness. A major interpretive problem here is that not only do we not know anything about the religious orientation of those for whom we have missing data, but we also have some concerns about those for whom we do have data. Suppose, for example, that religious liberals were far more likely to refuse to identify their religious preferences than the other two groups. What does that say about the data from those liberals who *do* self-identify? Do we actually have a distorted sample of liberals, or are the ones who didn't self-report just a random sample of liberals. For a more complete discussion of this issue, see Cohen, Cohen, West, and Aiken (2003).

Missing data on the dependent variable

We have somewhat different problems when data are missing on the dependent variable. When we have a design that reduces to a one-way analysis of variance or a *t* test, the treatment of missing data on the dependent variable is usually straightforward if we can assume that the data are at least MAR. Any software solution for an analysis of variance or *t* will provide a satisfactory result. The most serious problems we have are that our parameter estimates are better for large groups than for small ones. This assumes, of course, that our missingness is MAR and therefore ignorable. I don't mean to suggest that missing data are harmless in this situation, but the problem is more one of statistical power than interpretation.

But what about those situations where we would not be willing to assume that data are MAR, and therefore that the missingness mechanism is nonignorable? There are certainly situations where nonignorable missingness arises and creates problems. Imagine that we are running a treatment study for hypertension and people who are not receiving much benefit from the treatment start dropping out. Here missingness falls in the category of nonignorable. We will probably see that, for those remaining in our study, average blood pressure falls, but that may simply mean that we no longer have those unsuccessfully treated patients remaining in the study and raising the mean. All we have are data from those who remain, which largely means from those who derive benefit. In this case means and standard errors are going to be decidedly biased with respect to the parameters in the population, and we will be hard pressed to draw meaningful conclusions.

When it comes to designs that lead to a factorial analysis of variance, missing data is more of a problem. But even here the solutions are at least well spelled out, even if there is not always complete agreement on which solution is best.

It is easy to illustrate the problem caused by missing data in a factorial design. When we have a factorial with equal number of observations in each cell, then the main effects and interaction(s) are orthogonal to one another. Each effect is estimated independent of the others. We don't have to draw any conclusion conditional upon the level of another independent variable. However when we have unequal sample sizes, row, column, and interaction effects are confounded. As a simple, though extreme, example, consider the following design. In this experiment with hypothetical data we recorded data on driving errors from participants who had, and had not been drinking. We further broke the data down into those collected in Michigan and Arizona.

Illustration of the contaminating effects of unequal sample sizes

	Non-Drinking	Drinking	Row Means
Michigan	13 15 14 16 12	18 20 22 19 21	
		23 17 18 22 20	$\bar{X}_{1.} = 18.0$
	$\bar{X}_{11} = 14$	$\bar{X}_{12} = 20$	
Arizona	13 15 18 14 10	24 25 17 16 18	
	12 16 17 15 10 14		$\bar{X}_{2.} = 15.9$
	$\bar{X}_{21} = 14$	$\bar{X}_{22} = 20$	
Col Means	$\bar{X}_{.1} = 14$	$\bar{X}_{.2} = 20$	

The most obvious, and expected, result is that drivers who have been drinking make far more errors than drivers who have not been drinking. That will probably surprise no one. But notice also that drivers from Michigan appear to make more errors than drivers from Arizona. Is that really true? Are drivers from Michigan really that bad? If you look at the non-drinking drivers you see that Michigan and Arizona both have means of 14. And if you look at drinking drivers, the two states both have means of 20. So when we control for drinking, in other words when the results are treated as conditional on drinking, there is no between state effect. The higher score in Michigan actually came from the fact that there were proportionally more drinking drivers in that sample, and they made more errors because they had been drinking.

The example of drinking and driving errors was intended to point up the fact that missing data can cause important problems even in a simple factorial design. How we treat these data depends on why data are missing. Perhaps the data were collected by two different groups of researchers working in conjunction. The ones in Michigan decided that they would rather have twice as many Drinking than Non-Drinking Drivers. The researchers in Arizona made just the opposite choice for some reason. Then missingness does not depend in any way on the variables in the study, and is ignorable. In this case we would most likely want to partial all other effects out of the effect in question. Thus we look at States after partialling Drinking and the State X Drinking interaction (which in this example would be 0). Similarly for Drinking and for the interaction. This is the solution which SPSS and SAS call the Type I solution. It is the default in that software and should be used unless there is a very specific reason to do something else.

However let's assume for the moment that there are just many more drinking drivers in Michigan than in Arizona (and I have absolutely no reason to think that is really the case). Then it may be meaningful to say that Michigan drivers, on average, make more errors than Arizona drivers. The apparent cause is the higher percentage of drunken drivers in Michigan, but whatever the cause, there are still more driving errors in that state. This points out the important fact that even with a nice neat tidy analysis of variance, determining why the data are missing is important both in selecting an appropriate analysis and in drawing meaningful conclusions. If I really did think that

there were a higher percentage of drinking drivers in Michigan, I would not want to partial the Drinking variable in calculating a main effect for State.

Repeated measures designs

Within the category of experimental research designs we have repeated measures designs where participants are measured repeatedly over time or trials. The nice feature of these designs is that very often if you don't have data for one trial for a particular participant, you probably don't have data for other trials. The only thing you can do there is drop the participant from the analysis. Assuming that nonresponse is at least MAR, your parameter estimates will remain unbiased.

In some repeated measures, or time series, designs that take place over a period of time, there may be a different kind of problem with missing data. For example if the study takes place over a year and participants move away, get sick, or just get tired of the experiment, you will have data for the first few trials but not for later trials. There is not a simple solution to this problem. Simply dropping those individuals from the study is one possibility, and it may be an acceptable one if the data are MAR. If the data are not missing at random, but the poorer performing participants tend to drop out, then deleting whole cases will lead to bias in our estimates.

One solution that is sometimes employed, more often in medical research than in the social sciences, is called **Last Observation Carried Forward (LOCF)**. As the name implies, the last observation a participant gave is entered into the empty cells that follow (and hopefully the degrees of freedom are adjusted accordingly). In the past the FDA recommended this approach in clinical trials, but we now know that it leads to biased results and underestimates variability across trials. Similar strategies involve replacing missing observations with the participant's mean over the trials on which data are present, or basing imputed values on trends from past trials. All of these approaches carry with them assumptions about what the data would have looked like if the participant had not dropped out, and none of them is to be recommended. Methods discussed later in this article offer somewhat better solutions with less bias.

The intention-to-treat model

A common procedure in medical research, which is far less often used in the behavioral sciences, but which does have much to offer in many behavioral studies, is known as the intention-to-treat model. While it is not always thought of as a technique for missing data, that is exactly what it is because some number of participants in one condition are actually "missing for that condition" because they were switched to a different treatment.. (Gerard Dallal has a good web page on this topic at http://www.tufts.edu/~gdallal/itt.htm .)

Assume that we are doing a clinical study of two different treatments for angina. (I use a clearly medical example because I have useful data for that, but you could just as easily think of this study as a comparison of cognitive behavior therapy and family therapy as treatments for anorexia.) Assume further that patients were randomly assigned to a surgical or a medical treatment of angina. Two years later we record the number of patients who are still alive and who have died.

This sounds like a perfectly reasonable experimental design and we would expect a clear answer about which approach is best. But our patients are actual human beings, and the physicians who treat them have an ethical obligation to provide the best care possible. So although a patient is randomized to the medical treatment group, his physician may decide part way through the study that he really needs surgery. So what do we do with this patient? One approach would be to drop him from the study on the grounds that the randomized treatment assignment was not followed. Another possibility would be to reassign that patient to the surgical group and analyze his data "As-Treated." The third way would be to continue to regard him as being in the medical group regardless of what actually happened. This is the intention-to-treat model and, at first, it sounds crazy. We know the guy had surgery, but we pretend that he received medical treatment.

The first thing to recognize is that under the intention-to-treat model a null difference between groups must *not* be taken to mean that the two therapies are equivalent. As originally proposed by Richard Peto in the early 1980's, that was clearly part of the model, though it often gets forgotten. This is especially troublesome as "equivalence testing" is becoming more important in clinical settings. Suppose that we imagine that the Medical group was treated with a daily dose of castor oil. (I am of the generation that still remembers that wonderful stuff.) I would assume, though I am not a physician, that castor oil won't do anything for angina. The only thing it does is taste awful. After a short period the doctors of those in the castor oil group decide that it is a useless therapy and move most of their patients to the surgical group, which they have some ethical responsibility to do. So what has happened is that almost all of the patients were actually treated surgically, and, because they were treated alike, we would expect that they would respond alike. So when we run our statistical test at the end of treatment we would not be able to reject the null hypothesis. This certainly should not be taken to mean that castor oil is as good as surgery. We know that it clearly doesn't mean that. It simply says that if you assign some people to castor oil and some to surgery, they will all come out the same at the end. However, if the surgery group does come out with a significantly greater survival rate than the castor oil group, we have evidence that surgery is better than castor oil. So a statistically significant difference here means something, but a nonsignificant difference is largely uninterpretable. (Of course this was Fisher's model all along, but we often lose sight of that.)

In addition to analyzing the data as intent-to-treat, there is another analysis that we should be doing here. We should simply count the number of patients who ended up receiving which kind of treatment. When we discover that almost all patients were switched away from castor oil, that tells us a lot about what their physicians thought of the castor oil

treatment. It may also be very profitable to also run an analysis on groups "as-treated" and to present that result as well as the intent-to-treat result.

The following table shows the results of a study by the European Coronary Surgery Study Group, reported by Hollis and Campbell (1999), on surgical and medical treatment for angina pectoris. In that study 769 men were randomized to the two groups, 373 to the Medical treatment and 376 to the surgical treatment.

Results from Hollis and Campbell(1999)

	As-Assigned		**As-Treated**	
	Medical	Surgical	Medical	Surgical
Survivors	344	373	316	401
Deaths	29	21	33	17
Total	373	394	349	418
Mortality%.	7.8%	5.3%	9.5%	4.1%

We can see from the table that the As-Treated analysis would suggest that the surgery condition has a much lower mortality rate than medical condtion. However there were six patients who were assigned to surgery but died before that surgery could be performed, and were actually only treated medically. In the As-Treated analysis those six deaths raise the death rate for the medical group. In the Intent-To-Treat analysis we see that there are much smaller, and nonsignificant, differences between groups.

Contingency tables

Missing data is also a problem when the data are collected in the form of contingency tables, as they were in the intent-to-treat example above. Here we often cross our fingers and hope that the data are at least MAR. If they are not MAR, the interpretation of the results is cloudy at best. Here again the problems are the same ones that we have been discussing. If there is systematic dropout from one or more cells the missing data mechanism is confounded with the results of the data that are there.

We will take as an example a study by Gross (1985). She investigated attitude about weight in African-American and White high school girls. She sorted by ethnic group and recorded whether the girls wanted to gain weight, lose weight, or maintain their current weight. The data below are from her study, though she did not have any missing data. I have added the missing data to create an example.

Cohen et al. (2003) discuss the analysis of categorical data in detail and describe an imputation method that assigns missing data to the nonmissing cells of the table on the basis of a reasonable model of missingness. This is conceptually a very simple procedure. If we look at the African-American row in the following table we see that there are 30

missing observations—cases in which we know the ethnic group, but not their goals about weight gain or loss. But we also know that 24/99 = 24% of the African-American cases for which we *did* know the goal fell in the Gain column. So it seems reasonable to assume that 24% of the 30 missing cases would also have fallen in that column if we had been able to collect their data on Goal. Similarly, 24/55 = 44% of the group that wanted to gain weight were African-American, and so it is reasonable that 44% of the 10 missing cases in that column should be assigned to African-Americans. Therefore our new estimate of the count in the African-American/Gain cell should be 24 + (24/99)*30 = (24/55)*10 = 35.63. If we do the same for the rest of the cells we find the values indicated in parentheses in each cell.

	Goal				
Ethnic Group	**Gain**	**Lose**	**Maintain**	**Missing**	**Total Nonmissing**
African-American	24 (35.63)	47 (63.60)	28 (41.15)	30	99
White	31 (40.12)	352 (409.12)	152 (194.38)	60	535
Missing	10	20	30		
Total NonMissing	55	399	180		634

At this point we have allocated a total of 140.38 cases to the first row. Of those, 35.63/140.38 = 25.38% are in the African-American/Gain cell (whereas we formerly had 24/99 = 24.24% in that cell. In other words we have slightly changed our estimate of the percentage of observations falling in $cell_{11}$. Cohen et al. suggest reallocating the 30 missing observations in row 1 on the basis of this revised estimate, and performing similar calculations on each cell. If you do this you will again change, slightly, the percentage of observations in each cell. So you again reassign missing observations on the basis of those revised estimates. Eventually this iterative process will stabilize, with no further changes as a result of reallocation. At that point we declare the process completed, and run a standard chi-square test of that revised contingency table. For these data it took eight iterations for this process to stabilize when I did the calculations, and the resulting observed frequencies are shown below.

Observed frequencies after iteratively reallocating missing observations

Ethnic Group	Goal			Total
	Gain	Lose	Maintain	
African-American	36.49	63.01	42.30	141.80
White	39.96	407.38	194.86	642.20
Total	76.45	470.39	237.15	

The Pearson chi-square value for this table is 51.13 on 2 *df* which is clearly significant. For Gross's data (recall that she did not have the extra observations that I added via missing values), the chi-square was 37.23.

Observational studies

A high percentage of the research studies reported in the literature are nonexperimental. Those include standard regression studies, many studies of structural equation models, and survey studies among others. These studies do not use random assignment and are often limited to those who happen to fall within the sample at hand. Here there is even more opportunity for missing data, and perhaps even less chance of adequately modeling missingness. Moreover missing values are nearly as likely to occur with the independent variable (if there is one) as the dependent variable. Many methods have been developed to handle missingness in these situations, and the remainder of this chapter will focus on those. But keep in mind that these methods apply only when the data are at least missing at random.

Linear regression models

Many of our problems, as well as many of the solutions that have been suggested, refer to designs that can roughly be characterized as linear regression models. The problems, and the solutions, are certainly not restricted to linear regression—they apply to logistic regression, classification analyses, and other methods that rely on the linear model. But I will discuss the problem under the heading of linear regression because that is where it is most easily seen.

Suppose that we have collected data on several variables. One or more of those variables is likely to be considered a dependent variable, and the others are predictor, or independent, variables. We want to fit a model of the general form

$$Y_{ij} = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_{ij}$$

In this model data could be missing on any variable, and we need to find some way of dealing with that situation. We will assume that the missing data are either MCAR or MAR. A number of approaches to missingness in this kind of situation have been used over the years.

Casewise deletion

Probably the most common approach to missing data in regression analyses is what is called casewise deletion (or "listwise deletion," or "available case analysis"). Using this approach we simply drop from the analysis all cases that include any missing observation. The analysis is then carried out on the data that remain. This is usually the default analysis for most statistical software.

There are definite advantages to casewise deletion. If the missing data are at least MAR, casewise deletion leads to parameter estimates that are unbiased. The only loss is to statistical power, and in many situations this is not a particularly important consideration because this type of study often has a high level of power to begin with..

If the data are MNAR, this approach produces biased estimates. The resulting model is difficult to interpret because of confounding with missingness. However in many situations this approach has much to recommend it. It is certainly better than many of the alternatives.

Pairwise deletion

In pairwise deletion, data are kept or deleted on the basis of pairs of scores. In computing the overall covariance or correlation matrix, a pair of scores contributes to the correlation if both scores are present, but does not contribute if one or both of them are missing. Thus if a participant has data on Y, X_1, X_2, and X_5, but not on X_3 or X_4, that participant would be included in computing r_{YX1}, r_{YX2}, *and* r_{YX5} but not in computing r_{YX3} *or* r_{YX4} (and similarly for the rest of the pairs of observations). All available observations would be used in estimating means and standard deviations of the variables.

This method has one advantage, which is that it makes use of all available data and thus estimates parameters on the maximum sample size. But that is its only advantage. The major disadvantage is that each correlation, mean, and standard deviation is estimated on asomewhat different data set. In addition, it is not only possible, but not uncommon, that the covariance or correlation matrices resulting from this approach and needed for the analysis will not be positive definite. This means that it is impossible to calculate a normal inverse of either matrix, and thus not be able solve the necessary equations.

Pairwise deletion is generally a bad idea and I can think of no situation in which I would recommend it. As someone once said of stepwise regression, I would characterize pairwise deletion as "unwise" deletion.

Mean substitution

One approach that is sometimes taken when data on an independent variable are missing is to substitute for the missing scores the mean on that variable for all nonmissing cases. This approach has the dubious advantage of using all of the cases, but it has several disadvantages.

The following results were obtained using a data file from Cohen et al. (2003). In this situation we were predicting the Salary of members of the university faculty solely on the basis of the number of times their publications were cited (Citations). There are 62 cases with complete data and another 7 cases with Salary but without Citation. The results for an analysis of complete cases ($N = 62$) and an analysis of all cases with mean substitution for missing data ($N = 69$) are shown in the first two rows of the following table. Ignore the last row of the table for a moment.

Analysis	***N***	***r***	b_1	*St. error* (b_1)
Complete cases	62	.55	310.747	60.95
Mean substitution	69	.54	310.747	59.56
Mean substitution plus Missingness	69	.56	310.747	59.13

In this table you should notice that the regression coefficient for citations (b_1) is the same in the two analyses. However the standard error of the coefficient is smaller in the mean substitution analysis. This is because we have added seven cases where the deviation of the observation from the mean is 0, but we have increased the sample size. By holding the numerator constant, but increasing the denominator, automatically reduce the result. Although we have added cases, we have added no new information, and any change is in some way spurious. What we have is a standard error that is biased downward, leading to inappropriate test on b_1 and incorrect confidence limits. This is one of the reasons why mean substitution is not a particularly good way to proceed when you have missing data. It has been argued that if you have only a few missing cases, the use of mean substitution will lead to only minor bias. But if you have only a few missing cases you also have very little to gain by finding a way to add those cases into the analysis. I suggest that you don't even consider mean substitution.

Missing data coding

One way to improve on the mean substitution approach is to make use of any information supplied by missingness. A good way to do this is to add a variable to the regression that is coded "1" if the observation is missing and "0" if the observation is present. We again use mean substitution for the missing data.

Cohen was once an advocate of this approach, but his enthusiasm seems to have cooled over the years. The result of using both mean substitution and coding for missingness is

shown in the bottom row of the preceding table. There you can see that the coefficient for Citations remains the same, but the standard error is still underestimated. The one advantage is that the coefficient for the missingness variable of 4439 (not shown) represents the difference in mean income between those who do, and do not, have missing data on Citations. This is useful information, but we didn't need a regression solution to find it.

When we have multiple independent variables, Jones (1996) has shown that coding for missingness can lead to bias in both the regression coefficients and their standard errors. He examined a somewhat less biased approach, but still found that wanting. Coding for missingness in conjunction with mean substitution has not been particularly successful, and is no longer to be recommended.

Regression substitution (Imputation by least squares)

One additional fairly simple approach to the treatment of missing data is to regress the variable that has missing observations on the other independent variables (or even variables not used in the study), thus producing a model for estimating the value of a missing observation. We then use our regression equation to impute (substitute) a value for that variable whenever an observation is missing.

When there is a strong relationship between the variable that has missing observations and other independent variables, regression substitution is thought to work reasonably well. Lynch (2003) has characterized it was perhaps the best of the simple solutions to missing data. However regression imputation will increase the correlations among items because some of the items will have been explicitly calculated as a linear function of other items. This will affect the regression coefficients that result from the analysis. The imputed values would be expected to have less error than if the values were not missing. Thus regression imputation is likely to underestimate the standard error of the regression coefficients by underestimating the variance in the imputed variable. But this leads to an alternative solution, which will be discussed later, in which we resolve this problem by deliberately adding random error to our imputed observation.

In computing regression imputations, a fairly new procedure in SPSS, known as missing value analysis, by default adds a bit of error to each observation. We will see this in more detail later, but it is an attempt to reduce the negative bias in the estimated standard errors. This additional error does not solve the problem, but it reduces it somewhat. Like most imputation procedures, regression imputation assumes missing values are MAR (but not necessarily MCAR). The regression method also assumes homogeneity of regression, meaning that the same model explains the data for the non-missing cases and for the missing cases. If this assumption is false, the imputed values may be quite different from what the values would be if we had been able to measure them.

Hot deck imputation

One of the earliest methods of imputing missing values is known as hot deck imputation. Scheuren (2005) provides an interesting glimpse of how hot deck procedures developed within the U. S. Census Bureau. In the 1950's people generally felt that they had an obligation to respond to government surveys, and the nonresponse rate was low. In an effort to deal with unit nonresponse (the case where all data from a participant are missing), data cards (yes, they did use Hollerith cards in those days) for respondents were duplicated, and nonresponders were replaced by a random draw from these duplicate cards. Thus if you were missing a respondent of a certain gender from a certain census track, a draw was made from the data of respondents of that gender residing in that census track. The method worked well when only a small amount of data were missing, and the variance properties of the method were understood. Hansen, Hurwitz, and Madow (1953),

If it was acceptable to substitute "pseudo-respondents" for missing respondents, it was not a big step to replace missing items (questions) with pseudo-items. Again, items were replaced by a random draw from records selected on the basis of values on appropriate covariates. As long as the amount of missing data was minimal, this procedure worked well and was well understood. Unfortunately, the response rate to any survey or census has fallen over the years, and as we replace more and more data, the properties of our estimators, particularly their standard errors, becomes a problem. Hot deck imputation is not common today, although it is apparently useful in some settings.

Expectation-Maximization (EM)

The two most important treatments of missing data in the recent literature are expectation/maximization (known as the EM algorithm), Dempster, Laird, and Rubin (1977), and multiple imputation (MI) (Rubin, 1978). These are not distinct models, and EM is often used as a starting point for MI. I will discuss the two in turn, though they tend to blend together.

EM is a maximum likelihood procedure that works with the relationship between the unknown parameters of the data model and the missing data. As Schaefer (1998) has noted, "If we knew the missing values, then estimating the model parameters would be straightforward. Similarly, if we knew the parameters of the data model, then it would be possible to obtain unbiased predictions for the missing values." This suggests an approach in which we first estimate the parameters, then estimate the missing values, then use the filled in data set to re-estimate the parameters, then use the re-estimated parameters to estimate missing values, and so on. When the process finally converges on stable estimates we stop iterating.

For many, perhaps even most, situations in which we are likely to use EM, we will assume a multivariate normal model. Under that model it is relatively easy to explain in general terms what the EM algorithm does. Suppose that we have a data set with five variables ($X_1 - X_5$), with missing data on each variable. The algorithm first performs a

straightforward regression imputation procedure where it imputes values of X_1, for example, from the other four variables, using the parameter estimates of means, variances, and covariances or correlations from the existing data. (It is not important whether it calculates those estimates using casewise or pairwise deletion, because we will ultimately come out in the same place in either event.) After imputing data for every missing observation in the data set, EM calculates a new set of parameter estimates. The estimated means are simply the means of the variables in the imputed data set. But recall that when I discussed regression imputation I pointed out that the data imputed with that procedure would underestimate the true variability in the data because there is no error associated with the imputed observations. EM corrects that problem by estimating variances and covariances that incorporate the residual variance from the regression. For example, assume that we impute values for missing data on X_1 from data on X_2, X_3 , and X_4. To find the estimated mean of X_1 we simply take the mean of that variable. But when we estimate the variance of that variable we replace $\Sigma\left(X_i - \bar{X}\right)^2$ with $\Sigma\left(X_i - \bar{X}\right)^2 + s^2_{1.234}$. Similarly for the covariances. This counteracts the tendency to underestimate variances and covariances in regression imputation. Now that we have a new set of parameter estimates, we repeat the imputation process to produce another set of data. From that new set we re-estimate our parameters, as above, and then impute yet another set of data. This process continues in an iterative fashion until the estimates converge.

EM has the advantage that it produces unbiased, or nearly unbiased, estimates of means, variances, and covariances. Another nice feature is that even if the assumption of a multivariate normal distribution of observations is in error, the algorithm seems to work remarkably well.

One of the original problems with EM was the lack of statistical software. That is no longer a problem. The statistical literature is filled with papers on the algorithm and a number of programs exist to do the calculations. A good source, particularly because it is free and easy to use, is a set of programs by Joseph Schaefer. He has developed four packages, but only NORM is available to run as a stand alone under Windows. The others, CAT, which handles categorical data, MIX, for mixed models, and PAN, for panel or cluster data, are available as S-Plus libraries. Unfortunately S-Plus is not simple to use. These programs are available from http://www.stat.psu.edu/~jls/misoftwa.html. I show printout from NORM below, and it is quite easy to use. The paper by Schafer and Olson (1998) listed in the references is an excellent introduction to the whole procedure. SPSS version 13 also includes a missing data procedure that will do EM. The results of that procedure closely match that of NORM, but in my experience the standard errors in the resulting regression are smaller than those produced by data imputed using NORM.

An Example

The following example is based on data from a study by Compas (1990, personal communication) on the effect of parental cancer on behavior problems in children. The dependent variable is the Total Behavior Problem *T* score from the Achenbach Child Behavior Checklist (Achenbach, 1991). One might expect that the gender of the parent with cancer (SexP) would be a relevant predictor (things fall apart at home faster if mom

is sick than if dad is sick). Other likely predictors would the anxiety and depression scores of the cancer patient (AnxtP and DeptP) and the spouse (AnxtS and DeptS). These five predictors were to be used in a multiple linear regression analysis of behavior problems. Unfortunately, due to the timing of the first round of data collection, many of the observations were missing. Out of 89 cases, only 26 had complete data. The good thing is that it is reasonable to assume that missingness was due almost entirely to the timing of data collection (different families receive a diagnosis of cancer at different times) and not to the potential value of the missing values. So we can assume that the data are at least MAR without too much concern. The data for this example are available as an ASCII file and as an SPSS file at www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/.

Using only casewise deletion in SPSS, we obtain the following results. In the variable names "P" stands for "patient" and "S" for "spouse."

Casewise deletion using SPSS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-2.939	12.003		-.245	.809	-27.978	22.100
	sexp	-3.769	2.803	-.183	-1.344	.194	-9.616	2.079
	deptp	.888	.202	.764	4.393	.000	.467	1.310
	anxtp	-.064	.169	-.062	-.380	.708	-.417	.288
	depts	-.355	.155	-.460	-2.282	.034	-.679	-.030
	anxts	.608	.166	.719	3.662	.002	.262	.954

a. Dependent Variable: totbpt

$N = 26$, $R^2 = .658$

Notice that the sex of the parent with cancer does not have an effect, which is somewhat surprising, but that the patient's level of depression and the depression and anxiety levels of the spouse are all significant predictors. However, as noted above, complete data are only available for 26 of the 89 cases.

We can improve the situation using the EM algorithm as implemented by Schafer. An analysis of missingness on these variables is shown below.

Analysis of missing data

```
      NUMBER OF OBSERVATIONS =        89
      NUMBER OF VARIABLES =            9

                    NUMBER MISSING       %
MISSING
      Sexp                   7                 7.87
      deptp                 10                11.24
      anxtp                 10                11.24
      gsitp                 10                11.24
      depts                 29                32.58
      anxts                 29                32.58
      gsits                 29                32.58
      sexchild              48                53.93
      totbpt                48                53.93
```

Notice that for this analysis all of the variables in the data set are included. That will be true with imputation as well. In other words we will use variables in the imputation process that we may not use in the subsequent analysis, because those variables might be useful is predicting a participant's score, even if they are not useful in subsequently predicting behavior problems. This is especially important if you have variables that may be predictive of missingness.

The SPSS analysis of these EM-imputed data set follow.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-11.591	6.215		-1.865	.066	-23.953	.771
	SexP	-3.238	1.749	-.106	-1.851	.068	-6.717	.241
	DeptP	.886	.094	.722	9.433	.000	.699	1.073
	AnxtP	-.004	.099	-.003	-.039	.969	-.202	.194
	DeptS	-.418	.097	-.357	-4.310	.000	-.610	-.225
	AnxtS	.762	.099	.631	7.716	.000	.565	.958

a. Dependent Variable: TotBPt

$N = 89$ $R^2 = .871$

Notice that the regression coefficients are not drastically different from those in the previous analysis with casewise deletion, but the standard errors are considerably smaller. This is due mainly to the large increase in sample size with the imputed data. Interestingly the sex of the patient is much closer to significance at $\alpha = .05$. Notice also that the squared multiple correlation has increased dramatically, from .658 to .871. I am much more comfortable with this model than I was with the earlier one which was based on only 26 cases.

Multiple Imputation

One additional method for imputing values for missing observations is known as multiple imputation (MI). The original work on this approach was due to Rubin (1987), and it and EM are now becoming the dominant approaches to the treatment of missing data. A discussion of this material can be found in Allison (2001), Schafer & Olsen (1998), and Little (2005). There are a number of ways of performing MI, though they all involve the use of random components to overcome the problem of underestimation of standard errors. The parameter estimates using this approach are nearly unbiased.

The interesting thing about MI is that the word "multiple" refers not to the iterative nature of the process involved in imputation, but to the fact that we impute multiple complete data sets and run whatever analysis is appropriate on each data set in turn. We then combine the results of those multiple analyses using fairly simple rules put forth by Rubin (1987). In a way it is like running multiple replications of an experiment and then combining the results across the multiple analyses. But in the case of MI, the replications are repeated simulations of data sets based upon parameter estimates from the original study.

For many years the implementation of MI was held back by the lack of good algorithms by which to carry it out and by the lack of software. In the last 10 years or so both of those problems have been largely overcome. The introduction of new simulation methods known as Markov Chain Monte Carlo (MCMC) has simplified the task considerably, and software is now available to carry out the calculations. Schafer has implemented a method of Markov Chain Monte Carlo called data augmentation, and this approach is available in his NORM program referred to above. MI is not yet available in SPSS, but it is available in SAS as PROC MI and PROC MIANALYZE.

The process of multiple imputation, at least as carried out through data augmentation, involves two random processes. First, the imputed value contains a random component from a standard normal distribution. (I mentioned this in conjunction with the SPSS implementation of regression imputation.) Second, the parameter estimates used in imputing data are a random draw from a posterior probability distribution of the parameters.

The process of multiple imputation via data augmentation with a multivariate normal model is relatively straightforward, although I would hate to be the one who had to write the software. The first step involves the imputation of a complete set of data from parameter estimates derived from the incomplete data set. We could obtain these parameters directly from the incomplete data using casewise or pairwise deletion, or, as suggested by Schafer and Olsen (1998), we could first apply the EM algorithm and take our parameter estimates from the result of that procedure.

Under the multivariate normal model, the imputation of an observation is based on regressing a variable with missing data on the other variables in the data set. Assume, for simplicity, that X was regressed on only one other variable (Z). Denote the standard error

of the regression as $s_{X.Z}$. (In other words, $s_{X.Z}$ is the square root of $MS_{residual.}$). In standard regression imputation the imputed value of X ($\hat{X}$) would be obtained as

$$\hat{X}_i = b_0 + b_1 Z_i$$

But for data augmentation we will add random error to our prediction by setting

$$\hat{X}_i = b_0 + b_1 Z_i + u_i s_{X.Z}$$

where u_i is a random draw from a standard normal distribution. This introduces the necessary level of uncertainty into the imputed value. Following the imputation procedure just described, the imputed value will contain a random error component. Each time we impute data we will obtain a slightly different result.

But there is another random step to be considered. The process above treats the regression coefficients, and the standard error of regression as if they were parameters, when in fact they are sample estimates. But parameter estimates have their own distribution. (If you were to collect multiple data sets from the same population, the different analyses would produce different values of b_1, for example, and these estimates have a distribution.) So our second step will be to make a random draw of these estimates from their Bayesian posterior distributions—the distribution of the estimates given the data, or pseudodata, at hand.

Having derived imputed values for the missing observations, MI now iterates the solution, imputing values, deriving revised parameter estimates, imputing new values, and so on until the process stabilizes. At that point we have our parameter estimates and can write out the final imputed data file.

But we don't stop yet. Having generated an imputed data file, the procedure continues and generates several more data files. We do not need to generate many data sets, because Rubin has shown that in many cases three to five data sets are sufficient. Because of the randomness inherent in the algorithm, these data sets will differ somewhat from one another. In turn, when some standard data analysis procedure (here we are using multiple regression) is applied to each set of data, the results will differ slightly from one analysis to another. At this point we will derive our final set of estimates (in our case our final regression equation) by averaging over these estimates following a set of rules provided by Rubin.

I will illustrate the application of Rubin's method with the behavior problem example. I have used NORM to generate five imputed data sets, and have used SPSS to run the multiple regression of Total Behavior Problems on the five independent variables used previously. These regression coefficients and their squared standard errors for the five separate analyses are shown in the following table.

Regression coefficients from five imputed data sets

Data set	Estimated parameter	b_0	b_1	b_2	b_3	b_4	b_5
1	Coefficient	-11.535	-2.780	1.029	-.031	-0.359	0.572
	Variance	43.204	3.323	0.013	0.013	0.013	0.012
2	Coefficient	-11.501	-4.149	1.040	-0.093	-0.583	0.876
	Variance	40.488	2.680	0.010	0.009	0.009	0.007
3	Coefficient	-10.141	-5.038	0.766	0.123	-0.252	0.625
	Variance.	42.055	3.301	0.010	0.010	0.010	0.009
4	Coefficient	-11.533	-6.920	0.870	0.084	-0.458	0.815
	Variance	28.751	1.796	0.081	0.007	0.007	0.007
5	Coefficient	-14.586	-1.115	0.718	0.050	-0.373	0.814
	Variance	32.856	2.362	0.009	0.009	0.009	0.008
	Mean b_i	-11.859	-4.000	0.885	0.027	-0.405	0.740
	Mean Var.($\bar{W}$)	37.471	2.692	0.025	0.010	0.010	0.009
	Var. of b_i (B)	2.682	4.859	0.022	0.008	0.015	0.018
	T						
	$\sqrt{T}$	40.69	8.523	0.051	0.020	0.028	0.031
		6.379	2.919	0.226	0.141	0.167	0.176
	t	-1.859	-1.370	3.916*	0.191	2.425*	4.204*

* $p < .05$ "Var." refers to the squared standard error of the coefficient.

The final estimated regression coefficients are simply the means of the individual coefficients. Therefore

$$\hat{Y} = -11.859 - 4.000X_1 + 0.855X_2 + 0.027X_3 - 0.405X_4 + 0.740X_5$$

It is interesting to compare this solution with the solution from the analysis using casewise deletion. In that case

$$\hat{Y} = -2.939 - 3.769X_1 + 0.888X_2 - 0.064X_3 - 0.355X_4 + 0.608X_5$$

The variance of our estimates is composed of two parts. One part is the average of the variances in each column. These are shown labeled "Mean Var." in the last row. The other part is based on the variances of the estimated b_i. This is shown in the bottom row labeled "Var. of b_i." We then define $\bar{W}$ as the mean of the variances, and B as the variance of the coefficients. Then the Total variance of each estimated coefficient is

$$T = \bar{W} + \left(1 + \frac{1}{m}\right)B$$

The values of T and the square root of T, which is the standard error of the coefficient, are shown in the last row of the table. Below them is the result of $t = b_i / \sqrt{T}$, which is a t test

on the coefficient. Here you can see that the coefficients for the patient's depression score, the spouses depression score, and the spouses anxiety score are statistically significant at $\alpha = .05$.

Summary

This chapter has discussed many ways of dealing with missing data. In all cases missing data is a problem, but as we learn more about how to handle it, the problems become somewhat less important.

I pointed out that with standard experimental studies the solutions are relatively straightforward and don't lead to significant bias. In those studies missing data can lead to difficulty in interpretation, as shown in the example of driving performance under the influence of alcohol. But those problems are not going to be solved by any mathematical approach to missing data because they are at heart problems of logic rather than problems of mathematics.

With observational studies there are many methods that have been identified for dealing with missing observations. Some of the earlier solutions, such as hot deck imputation, mean substitution, and pairwise deletion are slowly tending to fall by the wayside because they lead to bias in parameter estimation. The most important techniques, now that the necessary software is available, are the expectation/maximization (EM) algorithm and multiple imputation (MI). Both of these rely on iterative solutions in which the parameter estimates lead to imputed values, which in turn change the parameter estimates, and so on. MI is an interesting approach because it uses randomized techniques to do its imputation, and then relies on multiple imputed data sets for the analysis. It is likely that MI will be the solution of choice for the next few years until something even better comes along.

References

Achenbach, T.M. (1991) *Manual for the Child Behavior Checklist/4 – 18 and 1991 profile*. Burlington, VT: University of Vermont Department of Psychiatry.

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

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003) *Applied multiple regression/correlation analysis for the behavioral sciences (3rd edition)*Mahwah, N.J.: Lawrence Erlbaum.

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

Gross, J.S. (1985) Weight modification and eating disorders in adolescent boys and girls. Unpublished doctoral dissertation, University of Vermont.

Hansen, M.H., Hurwitz, W, & Madow, W. (1953) *Sample survey methods and theory*. New York: Wiley

Hollis, S & Campbell, F. (1999) What is meant by intention to treat analysis? Survey of published randomised controlled trials. *British Journal of Medicine, 319*, 670-674.

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

Little, R. J. A. (1998) 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. (2005) Missing Data. In Everitt, B.S. & Howell, D.C. Encyclopedia of statistics in behavioral science. Chichester, Engand: Wiley

Little, R. J. A. and D. B. Rubin (1987). *Statistical analysis with missing data*. John Wiley & Sons, New York.

Lynch, S.M. (2003) Missing data. Available at http://www.princeton.edu/~slynch/missingdata.pdf .

Pickles, Andrew (2005). Missing data, problems and solutions. Pp. 689-694 in Kimberly Kempf-Leonard, ed., *Encyclopedia of social measurement*. Amsterdam: Elsevier.

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

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

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.* Chapman & Hall, London. Book No. 72, Chapman & Hall series Monographs on Statistics and Applied Probability.

Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*. 8: 3-15.

Schafer, J.L. and M. K. Olsen (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*. 33: 545-571.

Scheuren, F. (2005) Multiple imputation: How it began and continues. *The American Statistician, 59*, 315 – 319.

Sethi, S. & Seligman, M.E.P. (1993). Optimism and fundamentalism. *Psychological Science, 4*, 256-259.