TREATMENT OF MISSING DATA

David C. Howell

Missing data are a part of almost all research, and we all have to decide how to deal with it from time to time. There are a number of alternative ways of dealing with missing data, and this document is an attempt to outline those approaches. The original version of this document spent considerable space on using dummy variables to code for missing observations. That idea was popularized in the behavioral sciences by Cohen and Cohen (1983). However, that approach does not produce unbiased parameter estimates (Jones, 1996), and is no longer to be recommended--especially in light of the availability of excellent software to handle other approaches. For a very thorough book-length treatment of the issue of missing data, I recommend Little and Rubin (1987) .A shorter treatment can be found in Allison (2002) . I have recently written a chapter on missing data for an edited volume (Howell, 2007). Part of that paper forms the basis for some of what is found here. You can write me at David.Howell@uvm.edu for a copy of that missing data chapter. The current page will focus on analyses falling generally under the heading of multiple linear regression, and ignore some very real issues that arise from missing data in the analysis of variance and in the treatment of contingency tables. However if your interest is in missing data in a repeated measures ANOVA, you will find useful material at http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Mixed Models for Repeated Measures.pdf .

1.1 The nature of missing data

Missing completely at random

There are several reasons why the data may be missing. They may be missing because equipment malfunctioned, the weather was terrible, or people got sick, or the data were not entered correctly. Here the data are **missing completely at random (MCAR)**. When we say that data are missing completely at random, we mean that the probability that an observation (X_i) is missing is unrelated to the value of X_i or to the value of any other variables. Thus data on family income would *not* be considered MCAR if people with low incomes were less likely to report their family income than people with higher incomes. Similarly, if Whites were more likely to omit reporting income than African Americans, we again would not have data that were MCAR because missingness would be correlated with ethnicity. However if a participant's data were missing because he was stopped for a traffic violation and missed the data collection session, his data would presumably be missing completely at random. Another way to think of MCAR is to note that in that case any piece of data is just as likely to be missing as any other piece of data.

Notice that it is the value of the observation, and not its "missingness," that is important. If people who refused to report personal income were also likely to refuse to report family income, the data could still be considered MCAR, so long as neither of these had any relation to the income value itself. This is an important consideration, because when a data set consists of responses to several survey instruments, someone who did not complete the Beck Depression Inventory would be missing all BDI subscores, but that would not affect whether the data can be classed as MCAR.

This nice feature of data that are MCAR is that the analysis remains unbiased. We may lose power for our design, but the estimated parameters are not biased by the absence of data.

Missing at random

Often data are not missing *completely* at random, but they may be classifiable as **missing at random (MAR)**. For data to be missing *completely* at random, the probability that X_i is missing is unrelated to the value of X_i or other variables in the analysis. But the data can be considered as missing at random if the data meet the requirement that missingness does not depend on the value of X_i *after controlling for another variable.* For example, people who are depressed might be less inclined to report their income, and thus reported income will be related to depression. Depressed people might also have a lower income in general, and thus when we have a high rate of missing data among depressed individuals, the existing mean income might be lower than it would be without missing data. However, if, within depressed patients the probability of reported income was unrelated to income level, then the data would be considered MAR, though not MCAR.

The phraseology is a bit awkward here because we tend to think of randomness as not producing bias, and thus might well think that Missing at Random is not a problem. Unfortunately is is a problem, although in this case we have ways of dealing with the issue so as to produce meaningful and relatively unbiased estimates. But just because a variable is MAR does not mean that you can just forget about the problem.

Missing Not at random

If data are not missing at random or completely at random then they are classed as **Missing Not at Random (MNAR).** For example, if we are studying mental health and people who have been diagnosed as depressed are less likely than others to report their mental status, the data are not missing at random. Clearly the mean mental status score for the available data will not be an unbiased estimate of the mean that we would have obtained with complete data. The same thing happens when people with low income are less likely to report their income on a data collection form.

When we have data that are MNAR we have a problem. The only way to obtain an unbiased estimate of parameters is to model missingness. In other words we would need to write a model that accounts for the missing data. That model could then be incorporated into a more complex model for estimating missing values. This is not a task anyone would take on lightly. See Dunning and Freedman (2008) for an example.

1.2 The simplest approach--listwise deletion.

By far the most common approach is to simply omit those cases with missing data and to run our analyses on what remains. Thus if 5 subjects in group one don't show up to be tested, that group is 5 observations short. Or if 5 individuals have missing scores on one or more variables, we simply omit those individuals from the analysis. This approach is usually called listwise deletion, but it is also known as complete case analysis.

Although listwise deletion often results in a substantial decrease in the sample size available for the analysis, it does have important advantages. In particular, under the assumption that data are missing completely at random, it leads to unbiased parameter estimates. Unfortunately, even when the data are MCAR there is a loss in power using this approach. And when the data are not MCAR, bias results. (For example when low income individuals are less likely to report their income level, the resulting mean is biased in favor of higher incomes.) The alternative approaches discussed below should be considered in as a replacement for listwise deletion, though in some cases we may be better off to "bite the bullet" and fall back on listwise deletion.

1.3 A poor approach--pairwise deletion

Many computer packages offer the option of using what is generally known as pairwise deletion but has also been called "unwise" deletion. Under this approach each element of the intercorrelation matrix is estimated using all available data. If one participant reports his income and life satisfaction index, but not his age, he is included in the correlation of income and life satisfaction, but not in the correlations involving age. The problem with this approach is that the parameters of the model will be based on different sets of data, with different sample sizes and different standard errors. It is even quite possible tc generate an intercorrelation matrix that is not positive definite, which is likely to bring your whole analysis to a stop.

It has been suggested that if there are only a few missing observations it doesn't hurt anything to use pairwise deletion. But I would argue that if there are only a few missing observations that it doesn't hurt much to toss those participants out and use complete cases. If there are many missing observations you can do considerable harm with either analysis. In both situations the approaches given below are generally preferable.

1.3 Traditional treatments for missing data

Regression Models versus ANOVA models

I am about to make the distinction between regression and ANOVA models. This may not be the distinction that others might make, but it makes sense for me. I am really trying to distinguish between those cases for which group membership is unknown and cases in which the substantive variables are unknown.

Missing Identification of Group Membership

I will begin with a discussion of an approach that probably won't seem very unusual. In experimental research we usually know which group a subject belongs to because we specifically assigned them to that group. Unless we somehow bungle our data, group membership is not a problem. But in much applied research we don't always know group assignments. For example, suppose that we wanted to study differences in optimism among different religious denominations. We could do as Sethi and Seligman (1993) did and hand out an optimism scale in churches and synagogues, in which case we have our subjects pretty well classified with respect to religious affiliation because we know where we recruited them. However we could instead simply hand out the optimism scale to many people on the street and ask them to check off their religious affiliation. Some people might check "None," which is a perfectly appropriate response. But others might think that their religious affiliation is not our business, and refuse to check anything, leaving us completely in the dark. I would be hard pressed to defend the idea that this is a random event across all religious categories, but perhaps it is. Certainly "no response" is not the same as a response of "none," and we wouldn't want to treat it as if it were.

The most obvious thing to do in this situation would be to drop all of those non-responders from the analysis, and to try to convince ourselves that these are data missing completely at random. (Even if we did convince ourselves, I doubt we would fool our readers.) But a better approach is to make use of the fact that non-response is itself a bit of data, and to put those subjects into a group of their own. We would then have a specific test on the null hypothesis that non-responders are no different from other subjects in terms of their optimism score. And once we establish the fact that this null hypothesis is reasonable (if we should) we can then go ahead and compare the rest of the groups with somewhat more confidence. On the other hand, if we find that the non-responders differ systematically from the others on optimism, then we need to take that into account in interpreting differences among the remaining groups.

Example

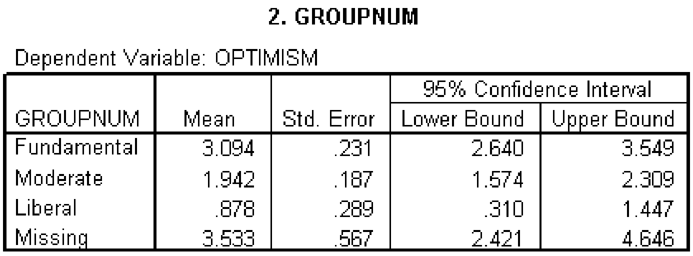
I will take the data from the study by Sethi and Seligman (1993) on optimism and religious fundamentalism as an example, although I will assume that data collection involved asking respondents to supply religious affiliation. These are data that I created and analyzed elsewhere to match the results that Sethi and Seligman obtained, although for purposes of this example I will alter those data so as to remove "Religious Affiliation" from 30 cases. I won't tell you whether I did this randomly or systematically, because the answer to that will be part of our analysis. The data for this example are contained in a file named FundMiss.dat, which is available for downloading, although it is much too long to show here. (The variables are, in order, ID, Group (string variable), Optimism, Group Number (a numerical coding of Group), Religious Influence, Religious Involvement, Religious Hope, and Miss (to be explained later).) You will note that when respondents are missing any data, the data are missing on Group membership and on all three religiosity variables, but not on Optimism. (Missing values are designated here with a period (.). If your software doesn't like periods as missing data, you can take any editor and change periods to asterisks (*), or blanks, or 999s, or whatever it does like.) This is the kind of result you might find if the religiosity variables all come off the same measurement instrument and that instrument also has a place to record religious affiliation. We see cases like this all the time. The dependent variable for these analyses is the respondent's score on the Optimism scale, and the resulting sample sizes, means, and standard deviations are shown in Table 1, as produced by SPSS.

```
            - - Description of Subpopulations - -
Summaries of      OPTIMISM
By levels of      GROUPNUM   Group Membership
Variable      Value  Label                   Mean    Std Dev     Cases
For Entire Population                      2.1633     3.2053       600
GROUPNUM          1  Fundamentalist        3.0944     2.8573       180
GROUPNUM          2  Moderate              1.9418     3.1629       275
GROUPNUM          3  Liberal                .8783     3.2985       115
GROUPNUM          4  Missing               3.5333     3.1919        30
  Total Cases = 600
```

Table 1 Descriptive Statistics for Optimism as a Function of Group Membership

From this table we see that there are substantial differences among the three groups for whom Religious Affiliation is known. We also see that the mean for the Missing subjects is much closer to the mean of Fundamentalist than to the other means, which might suggest that Fundamentalists were more likely to refuse to provide a religious affiliation than were members of the other groups.

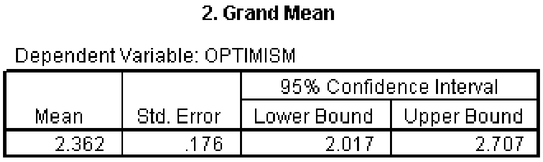
The results of an analysis of variance on Optimism scores of all four groups is presented in Table 2. Here I have asked SPSS to use what are called "Simple Contrasts" with the last (missing) group as the reference group. This will cause SPSS to print out a comparison of each of the first three groups with the Missing group. I chose to use simple contrasts because I wanted to see how Missing subjects compared to each of the three non-missing groups.



2. GROUPNUM

Dependent Variable: OPTIMISM

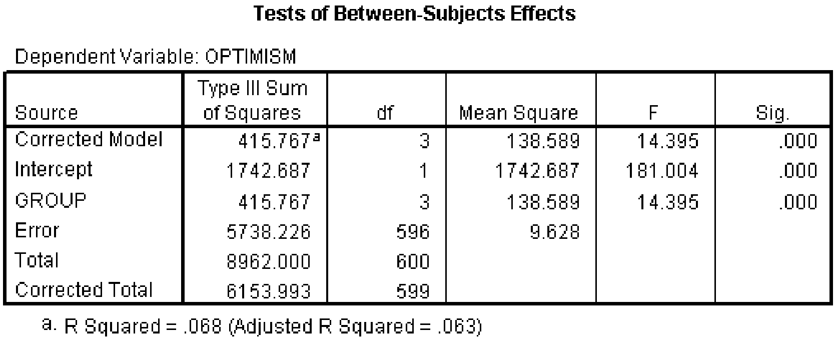
GROUPNUM	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Fundamental	3.094	.231	2.640	3.549
Moderate	1.942	.187	1.574	2.309
Liberal	.878	.289	.310	1.447
Missing	3.533	.567	2.421	4.646



2. Grand Mean

Dependent Variable: OPTIMISM

Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
2.362	.176	2.017	2.707

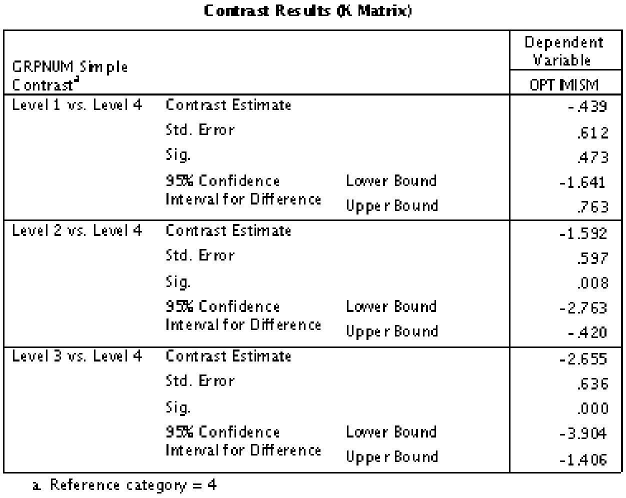


Tests of Between-Subjects Effects

Dependent Variable: OPTIMISM

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	415.767[a]	3	138.589	14.395	.000
Intercept	1742.687	1	1742.687	181.004	.000
GROUP	415.767	3	138.589	14.395	.000
Error	5738.226	596	9.628		
Total	8962.000	600			
Corrected Total	6153.993	599			

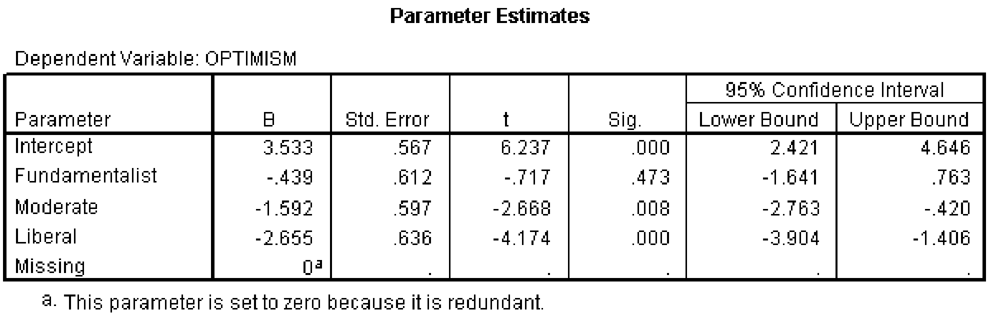
a. R Squared = .068 (Adjusted R Squared = .063)



Contrast Results (K Matrix)

GRPNUM Simple Contrast[a]			Dependent Variable
			OPTIMISM
Level 1 vs. Level 4	Contrast Estimate		-.439
	Std. Error		.612
	Sig.		.473
	95% Confidence Interval for Difference	Lower Bound	-1.641
		Upper Bound	.763
Level 2 vs. Level 4	Contrast Estimate		-1.592
	Std. Error		.597
	Sig.		.008
	95% Confidence Interval for Difference	Lower Bound	-2.763
		Upper Bound	-.420
Level 3 vs. Level 4	Contrast Estimate		-2.655
	Std. Error		.636
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-3.904
		Upper Bound	-1.406

a. Reference category = 4



Parameter Estimates

Dependent Variable: OPTIMISM

					95% Confidence Interval	
Parameter	B	Std. Error	t	Sig.	Lower Bound	Upper Bound
Intercept	3.533	.567	6.237	.000	2.421	4.646
Fundamentalist	-.439	.612	-.717	.473	-1.641	.763
Moderate	-1.592	.597	-2.668	.008	-2.763	-.420
Liberal	-2.655	.636	-4.174	.000	-3.904	-1.406
Missing	0[a]	.	.	.	.	.

[a] This parameter is set to zero because it is redundant.

Table 2 Analysis of Variance with All Four Groups -- Simple Contrasts

At the top of Table 2 you see the means of the four groups, as well as the unweighted mean of all groups (labeled Grand Mean). Next in the table is an Analysis of Variance summary table, showing that there are significant differences between groups; ($F_{3,599}$ = 14.395; *p* = .000).

A moment's calculation will show you that the difference between the mean of Fundamentalists and the mean of the Missing group is 3.094 - 3.533 = -0.439. Similarly the Moderate group mean differs from the mean of the Missing group by 1.942 - 3.533 = -1.591, and the Liberal and Missing means differ by 0.878 - 3.533 = -2.655. Thus participants who do not give their religious affiliation have Optimism scores that are much closer to those of Fundamentalists than those of the other affiliations.

In the section of the table labeled "Parameter Estimates" we see the coefficients of -.439, 1.592, and -2.655. You should note that these coefficients are equal to the difference between each group's mean and the mean of the last (Missing) group. Moreover, the *t* values in that section of the table represent a significance test on the deviations from the mean of the Missing group, and we can see that Missing deviates significantly from Moderates and Liberals, but not from Fundamentalists. This suggests to me that there is a systematic pattern of non- response which we must keep in mind when we evaluate our data. Subjects are not missing at random because missingness depends on the value of that variable. (Notice that the coefficient for missing is set at 0 and labeled "redundant." It is redundant because if someone is not in the Fundamentalist, Moderate, or Liberal group, we know that they are missing. "Missing," in this case, adds no new information.)

Orthogonal Contrasts

You might be inclined to suggest that the previous analysis doesn't give us exactly what we want because it does not tell us about relationships among the three groups having non-missing membership. In part, that's the point, because we wanted to include all of the data in a way that told us something about those people who failed to respond, as well as those who did supply the necessary information.

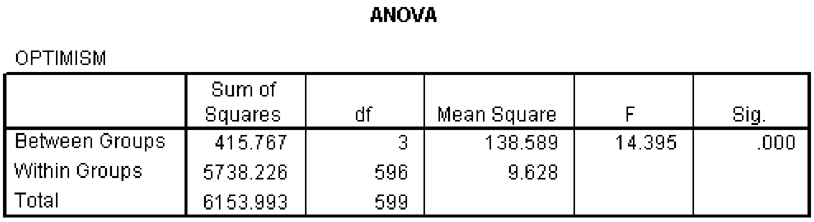
However, for those who want to focus on those subjects who provided Religious Affiliation while not totally ignoring those who did not, an alternative analysis would involve the use of orthogonal contrasts not only to compare the non-responders with all responders, but also to make specific comparisons among the three known groups. But keep in mind that because the data are not MCAR the means, particularly the grand mean, is likely to be biased. (If Fundamentalists are less likely to respond, and if they have higher optimism scores, the grand mean of optimism will be biased downward from what it would have been had they responded.)

You can use SPSS (OneWay) or any other program to perform the contrasts in question. (Or you can easily do it by hand). Suppose that I am particularly interested in knowing how the non-responders differ from the average of all responders, but that I am also interested in comparing the Moderates with the other two identified groups, and then the Fundamentalists with the Liberals. I can run these contrasts by providing SPSS with the following coefficients.

```
Missing vs Non-Missing                              1     1   1     -3
(Fundamental & Moderate) vs Liberals                1     1  -2      0
Fundamental vs Liberal                              1    -1   0      0
```

The first contrast deals with those missing responses that have caused us a problem, and the second and third contrasts deal with differences among the identified groups. The results of this analysis are presented below. (I have run this using SPSS syntax because it produces more useful printout.)

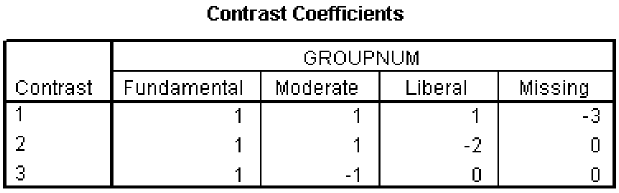
```
ONEWAY
   optimism BY groupnum(1 4)
   /CONTRAST= 1 1 1 -3  /CONTRAST= 1 1 -2  /CONTRAST= 1 -1
   /HARMONIC NONE
   /FORMAT NOLABELS
   /MISSING ANALYSIS .
```



ANOVA

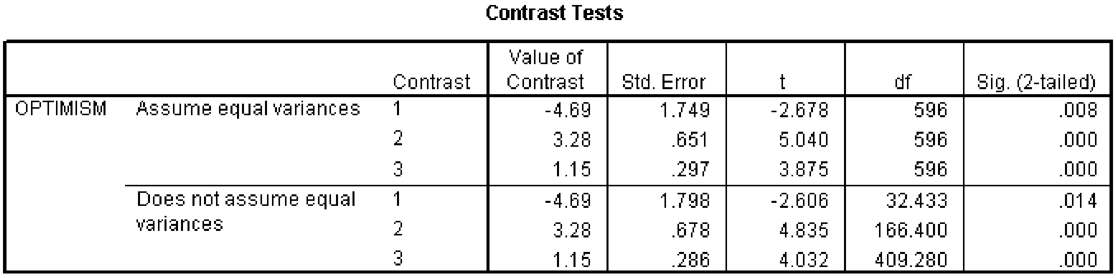
OPTIMISM

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	415.767	3	138.589	14.395	.000
Within Groups	5738.226	596	9.628		
Total	6153.993	599			



Contrast Coefficients

	GROUPNUM			
Contrast	Fundamental	Moderate	Liberal	Missing
1	1	1	1	-3
2	1	1	-2	0
3	1	-1	0	0



Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
OPTIMISM	Assume equal variances	1	-4.69	1.749	-2.678	596	.008
		2	3.28	.651	5.040	596	.000
		3	1.15	.297	3.875	596	.000
	Does not assume equal variances	1	-4.69	1.798	-2.606	32.433	.014
		2	3.28	.678	4.835	166.400	.000
		3	1.15	.286	4.032	409.280	.000

Table 3 OneWay Analysis of Variance on Optimism with Orthogonal Contrasts

Notice in Table 3 that the contrasts are computed with and without pooling of error terms. In our particular case the variances are sufficiently equal to allow us to pool error, but, in fact, for these data it would not make any important difference to the outcome which analysis we used. In Table 3 you will see that all of the contrasts are significant. This means that non-responders are significantly different from (and more optimistic than) responders, that Fundamentalists and Moderates combined are more Optimistic than Liberals, and that Fundamentalists are in turn more optimistic than Moderates.

I have presented this last analysis to make the point that you have not lost a thing by including the missing subjects in your analysis relative to running the analysis excluding missing observations. The second and third contrasts are exactly the same as you would have run if you had only used the three identified groups. However, this analysis includes the variability of Optimism scores from the Missing group in determining the error term, giving you somewhat more degrees of freedom. In a sense, you can have your cake and eat it too, although, as I noted above, the overall mean is biased relative to what it would have been had we collected complete data.

Cohen and Cohen (1983, Chapter 7) provide additional comments on the treatment of missing group membership, and you might look there for additional ideas. In particular, you might look at their treatment of the case where there is missing information on more than one independent variable.

This situation, where data on group membership is missing, is handled by the analysis above. Notice that, other than the overall mean, the analysis is not dependent on the nature of the mechanism behind missingness, which is in fact addressed by the analysis. This will not necessarily be the case in the following analysis, where the nature of missingness is important.

Missing Data on the Dependent Variable

We have a different kind of problem when we have data missing on the dependent variable that makes the results of our study much more difficult to interpret. If our data are in the form of a one-way analysis of variance, and if we can assume that data are missing completely at random, things are not particularly bad. We will lose power because of smaller sample sizes, and the means of larger groups will be estimated with less error than means of smaller groups, but we will not have problems with biased estimates. But keep in mind that I'm speaking here of data that are missing completely at random.

But suppose that our data are not missing completely at random. Suppose that we are comparing two treatments for hypertension. In the ideal study we have all participants take the medication they are prescribed and then we compare blood pressure levels at the end of treatment. But in the real world we know that there is usually a dropout problem in medical studies. In particular, those who are not helped by the treatment are more likely to drop out, or to die. If one drug is quite successful and the other is pretty much a failure, the sample size will be very much smaller in the second treatment. Moreover, those who remain, and whose blood pressure is eventually measured, are likely to be the ones who benefitted from the treatment. So if we see that the means of the two groups are nearly equal at the end of treatment, we might be led to the conclusion that the two treatments are equally effective. In fact, one was a horrible treatment but we didn't have data from its "failures." In such a setting missing data make the interpretation of means quite risky. (Perhaps the most appropriate statistic would be the drop out rate instead of the mean.)

1.4 Other Not-So-Good Approaches

I want to talk about a few approaches that are sometimes used and that we know are not very wise choices. It is important to talk about these because it is important to discourage their use, but more importantly because they lead logically to modern approaches that are very much better.

Hot deck imputation

Hot deck imputation goes back over 50 years and was used quite successfully by the Census Bureau and others in better times. In the 1940's and '50's most citizens seemed to feel that they had a responsibility to fill out surveys, and, as a result, relatively little data was missing. Suppose that in the 1950 census a young, black, male resident of census block 32a either was not available or refused to participate. The census bureau would simply take a stack of Hollerith cards (you may know them as "IBM cards") that came from young, black males in census block 32a, reach in the pile, and pull one out. That card was substituted for the missing card and the analysis continued. That is not as outrageous a procedure as it might seem at first glance. First of all there were relatively few missing observations to be replaced. Second the replacement data was a random draw from a collection of data on similar participants. Third the statistical implications of this process were thought to be pretty well understood. I don't believe that hot deck imputation is much used anymore, but it served its purpose at the time. Scheuren (2005) has an interesting discussion of how the process was developed within the U. S. Census Bureau.

Mean substitution

An old procedure that should certainly be relegated to the past was the idea of substituting a mean for the missing data. For example, if you don't know my systolic blood pressure, just substitute the mean systolic blood pressure for mine and continue. There are a couple of problems with this approach. In the first place it adds no new information. The overall mean, with or without replacing my missing data, will be the same. In addition, such a process leads to an underestimate of error. Cohen et al. (2003) gave an interesting example of a data set on university faculty. The data consisted of data on salary and citation level of publications. There were 62 cases with complete data and 7 cases for which the citation index was missing. Cohen gives the following table.

Analysis	**N**	**r**	**b**	**St. Err. b**
Complete cases	62	.55	310.747	60.95
Mean substitution	69	.54	310.747	59.13

Notice that using mean substitution makes only a trivial change in the correlation coefficient and no change in the regression coefficient. But the st. err (b) is noticeably smaller using mean substitution. That should not be surprising. We have really added no new information to the data but we have increased the sample size. The effect of increasing the sample size is to increase the denominator for computing the standard error, thus reducing the standard error. Adding no new information certainly should not make you more comfortable with the result, but this would seem to. The reduction is spurious and should be avoided--as we'll see below.

Regression substitution

If we don't like mean substitution, why not try using linear regression to predict what the missing score should be on the basis of other variables that are present? This approach has been around for a long time and has at least one advantage over mean substitution. At least the imputed value is in some way conditional on other information we have about the person. With mean substitution, if we are missing a person's weight we assign him the average weight. Put somewhat incorrectly, with regression substitution we would assign him the weight of males of around the same age. That has to be an improvement. But the problem of error variance remains. By substituting a value that is perfectly predictable from other variables, we have not really added more information but we have increased the sample size and reduced the standard error.

There is one way out of this difficulty, however, and SPSS has implemented it in their Missing Value Analysis procedure. By default that procedure adds a bit of random error to each substitution. That does not totally eliminate the problem, but it does reduce it. There are better ways, however, and they build on this simple idea.

1.5 Modern Approaches--Maximum Likelihood and Multiple Imputation

I am going to go fairly lightly on what follows, because the solutions are highly technical. I will, however, give references to good discussions of the issues and to software. As implied by the title of this section, the solutions fall into two categories--those which rely on maximum likelihood solutions, and those which involve multiple imputation.

Maximum Likelihood

The principle of maximum likelihood is fairly simple, but the actual solution is computationally complex. I will take an example of estimating a population mean, because that illustrates the principle without complicating the solution.

Suppose that we had the sample data 1, 4, 7, 9 and wanted to estimate the population mean. You probably already know that our best estimate of the population mean is the sample mean, but forget that bit of knowledge for the moment.

Suppose that we were willing to assume that the population was normally distributed, simply because this makes the argument easier. Let the population mean be represented by the symbol μ, although in most discussions of maximum likelihood we use a more generic symbol, θ, because it could stand for any parameter we wish to estimate.

We could calculate the probability of obtaining a 1, 4, 7, and 9 for a specific value of μ. This would be the product p(1)*p(4)*p(7)*p(9). You would probably guess that this probability would be very very small if the true value of μ = 10, but would be considerably higher if the true value of μ were 4 or 5. (In fact, it would be at its maximum for μ = 5.25.) For each different value of μ we could calculate p(1), etc. and thus the product. For some value of μ this product will be larger than for any other value of μ. We call this the maximum likelihood estimate of μ. It turns out that the maximum likelihood estimator of the population mean is the sample mean, because we are more likely to obtain a 1, 4, 7, and 9 if μ = the sample mean than if it equals any other value.

The same principle applies in regression, although it is considerably more complicated. If we assume a multivariate normal distribution, we can calculate maximum likelihood estimators for the means, variances, and covariance given the sample data. These are the values of those parameters that would make the data we obtained maximally likely. Once we have these estimates, we can use them to derive the optimal regression equation.

The EM Algorithm

There are a number of ways to obtain maximum likelihood estimators, and one of the most common is called the Expectation-Maximization algorithm, abbreviated as the EM algorithm. The basic idea is simple enough, but the calculation is more work than you would want to do.

Schafer (1999) phrased the problem well when he 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 obtained unbiased predictions for the missing values." Here we are going to do both. We will first estimate the parameters on the basis of the data we do have. Then we will estimate the missing data on the basis of those parameters. Then we will re-estimate the parameters based on the filled-in data, and so on. We would first take estimates of the variances, covariances and means, perhaps from listwise deletion. We would then use those estimates to solve for the regression coefficients, and then estimate missing data based on those regression coefficients. (For example, we would use whatever data we have to estimate the regression $\hat{Y} = bX + a$, and then use X to estimate Y wherever it is missing.) This is the estimation step of the algorithm. Having filled in missing data with these estimates, we would then use the complete data (including estimated values) to recalculate the regression coefficients. But recall that we have been worried about underestimating error in choosing our estimates. The EM algorithm gets around this by adding a bit of error to the variances it estimates, and then uses those new estimates to impute data, and so on until the solution stabilizes. At that point we have maximum likelihood estimates of the parameters, and we can use those to make the final maximum likelihood estimates of the regression coefficients.

There are alternative maximum likelihood estimators that will be better than the ones obtained by the EM algorithm, but they assume that we have an underlying model (usually the multivariate normal distribution) for the distribution of variables with missing data.

An excellent discussion of the EM algorithm and its solution is provided by Schafer (1997, 1999, and Schafer & Olsen (1998). Schafer has also provided a program that will do the imputation. That program is named NORM and is freely available from http://www.stat.psu.edu/~jls/misoftwa.html. The Missing Value Analysis module in SPSS version 13 also includes a missing data procedure that will do EM. In my experience the standard errors that SPSS produces are smaller than those in data imputed by NORM. A Web page that I wrote for one of our graduate students on using the stand alone NORM program can be found at NormMissingData.html

Multiple Imputation

An alternative the maximum likelihood method is called Multiple Imputation. Each of the solutions that we have discussed involves estimating what the missing values would be, and using those "imputed" values in the solution. With dummy variable coding we substituted a constant (often the variable mean) for the missing data. For the EM algorithm we substituted a predicted value on the basis of the variables that were available. In multiple imputation we will substitute random data.

In multiple imputation we generate imputed values on the basis of existing data, just as we did with the EM algorithm. But suppose that we are estimating Y on the basis of X. For every situation with X = 5, for example, we will impute the same value of Y. This leads to an underestimate of the standard error of our regression coefficients, because we have less variability in our imputed data than we would have had if those values had not been missing. One solution was the one used in the EM algorithm, where we altered the calculational formulae by adding in error in the calculation. With multiple imputation we are going to take our predicted values of Y and then add an error component drawn randomly from the residual distribution of $Y - \hat{Y}$. This is known as "random imputation."

This solution will still underestimate the standard errors. We solve this problem by repeating the imputation problem several times, generating multiple sets of new data whose coefficients vary from set to set. We then capture this variability and add it back into our estimates. This is a very messy process, and the reader is referred to Allison (2002), Little and Rubin (1987), and Schafer (1999) for the technical details.

One of the major problems with MI for years was the absence of simple software. New simulation methods known as Marcov Chain Monte Carlo (MCMC) has simplifiedthe task considerably and software is now available to carry out the calculations. As part of Schafer's NORM program discussed above, his data augmentation procedure will perform MI. As of this writing SPSS does not provide the capability of doing MI, but PROC MI and PROC MIANALYZE in SAS will do it.

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.

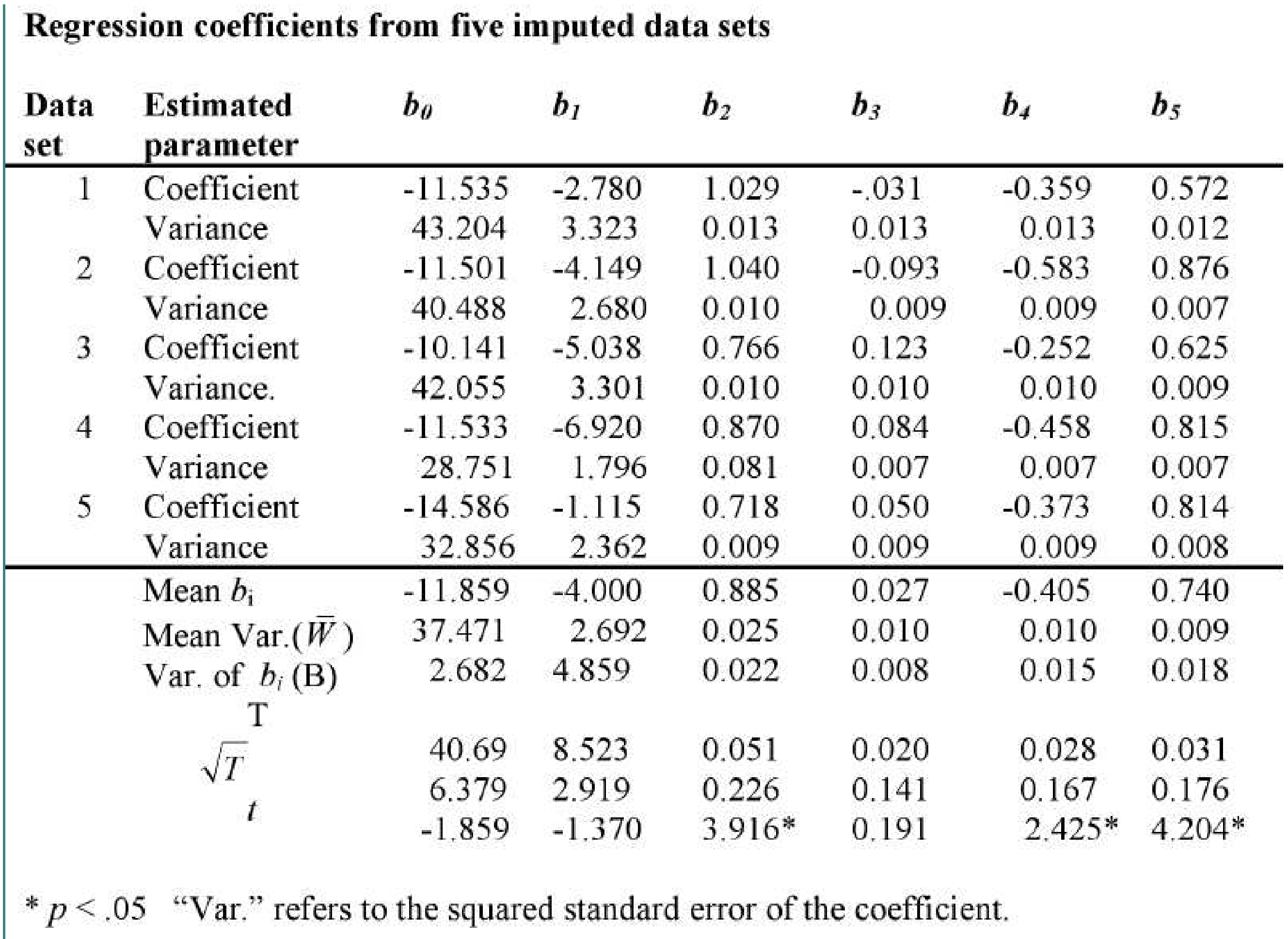
Under the multivariate normal model, the imputation of an observation is based on regressing a variable with missing data on the other variables in the data set. Assume, for simplicity, that X was regressed on only one other variable (Z). Denote the standard error of the regression as s $_{X.Z}$. (In other words, s$_{X.Z}$ is the square root of MSresidual.). In standard regression imputation the imputed value of X ($\bar{X}$) 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	40.69	8.523	0.051	0.020	0.028	0.031
	$\sqrt{T}$	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 α = .05.

Using the NORM Stand-Alone Program

I have spoken above about the stand-alone NORM program that Schafer has written. I gave the link to that program above. I was asked by a former student if I could write something that was a step-by-step approach to using NORM, and that document is available at "NormMissingData.html".

Using SAS software

The easiest way to accomplish what I have described above is to use SAS if you have it available. I have created a SAS program and a text file with the data. The file is named cancerdot.dat, where the "dot" indicates that the missing values are input as periods. You will have to modify the SAS program slightly to point to the correct file in the correct folder (directory), but that should not be a problem.

References

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

Cohen, J. & Cohen, P. (1983) Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.).Hillsdale, NJ: Erlbaum. Return

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. Return

Dunning, T., & Freedman, D.A. (2008) Modeling section iffects. in Outhwaite, W. & Turner, S. (eds) Handbook of Social Science Methodology. London: Sage. Return

Howell,D. C. (2007) The analysis of missing data. In Outhwaite, W. & Turner, S. Handbook of Social Science Methodology. London: Sage.Return

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

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. Return

Schafer, J. L. (1997). Analysis of incomplete multivariate data, London, Chapman & Hall.">

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

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

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

Return to Dave Howell's Statistical Home Page

Send mail to: David.Howell@uvm.edu)

Last revised 3/7/2009