Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data

Six: Dealing with Missing or Incomplete Data: Debunking the Myth of Emptiness

Contributors: Jason W. Osborne
Book Title: Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data
Chapter Title: "Six: Dealing with Missing or Incomplete Data: Debunking the Myth of Emptiness"
Pub. Date: 2013

Access Date: November 11, 2015
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412988018
Online ISBN: 9781452269948
DOI: http://dx.doi.org/10.4135/9781452269948.n6
Print pages: 105-139

http://dx.doi.org/10.4135/9781452269948.n6*Old Dominion University*

[p. 105 ↓]

Six: Dealing with Missing or Incomplete Data: Debunking the Myth of Emptiness

In almost any research you perform, there is the potential for missing or incomplete data. Missing data can occur for many reasons: participants can fail to respond to questions (legitimately or illegitimately—more on that later), equipment and data collecting or recording mechanisms can malfunction, subjects can withdraw from studies before they are completed, and data entry errors can occur. In later chapters I also discuss the elimination of extreme scores and outliers, which also can lead to missingness.

The issue with missingness is that nearly all classic and modern statistical techniques assume (or require) complete data, and most common statistical packages default to the least desirable options for dealing with missing data: deletion of the case from the analysis. Most people analyzing quantitative data allow the software to default to eliminating important data from their analyses, despite that individual or case potentially having a good deal of other data to contribute to the overall analysis.

It is my argument in this chapter that all researchers should examine their data for missingness, and researchers wanting the best (i.e., the most replicable and generalizable) results from their research need to be prepared to deal with missing data in the most appropriate and desirable way possible. In this chapter I briefly review common reasons for missing (or incomplete) data, compare and contrast several common methods for dealing with missingness, and demonstrate some of the benefits of using more modern methods (and some drawbacks of using the traditional, default methods) in the search for the best, most scientific outcomes for your research.

[p. 106 ↓]

Is emptiness meaninglessness?

Modern researchers seem to view missing data as empty, useless, a void that should have been filled with information, a thing without pattern, meaning, or value.

Yet the ancient Greeks saw potential in emptiness. The Greek goddess Chaos (Khaos) represented unfilled space (initially the unfilled space between the earth and the heavens in the creation mythology), much as a blank canvas represents unfilled potential to an artist or a blank page to a writer. And ancient Olmec, Indian, and Arabic mathematicians saw usefulness in the mathematical quantification of nothing, what we now call zero (Colebrooke, 1817; Diehl, 2004).

The modern computer era is built upon use of 0s and 1s as indicators of important states, both meaningful and critical to the functioning of devices that are now ubiquitous. Just as our ancestors saw usefulness and information in absence, I propose to demonstrate that missingness can not only be informative, but in certain circumstances can also be filled with meaning and that those with missing data do not need to be banished from our analyses but rather can contribute to a more complete and accurate understanding of the population about which we wish to draw conclusions.

What is Missing or Incomplete Data?

The issue before us is whether we have complete data from all research participants on all variables (at all possible time points, if it is a repeated-measures design). If any data on any variable from any participant is not present, the researcher is dealing with missing or incomplete data. For the purposes of the rest of this chapter, we use the term *missing* to indicate that state of affairs. In many types of research, it is the case that there can be *legitimate missing data*. This can come in many forms, for many reasons. Most commonly, legitimate missing data is an absence of data when it is appropriate for there to be an absence. Imagine you are filling out a survey that asks you whether you are married,[1] and if so, how long you have been married. If you say you are not married, it is legitimate for you to skip the follow-up question on how long you have been married. If a survey asks you whether you voted in the last election, and if so, what party the candidate was from, it is legitimate to skip the second part if you did not vote in the last election.

In medical research, it is possible that whatever treatment a participant is receiving has eliminated the condition that person was getting treated for (since I am not a medical doctor, I will call that “being cured”). In a long-term study of people receiving a particular type of treatment, if you are no longer receiving treatment because you are cured, **[p. 107 ↓]** that might be a legitimate form of missing data. Or perhaps you are following employee satisfaction at a company. If an employee leaves the company (and thus is no longer an employee) it seems to me legitimate that person should no longer be responding to employee satisfaction questionnaires.

Large data sets, especially government data sets, are full of legitimately missing data, and researchers need to be thoughtful about handling this issue appropriately (as I hope you will be thoughtful about all issues around data cleaning). Note too that even in the case of legitimate missingness, missing-ness is meaningful. Missingness in this context informs and reinforces the status of a particular individual and can even provide an opportunity for checking the validity of an individual's responses. In cleaning the data from a survey on adolescent health risk behaviors many years ago, I came across some individuals who indicated on one question that they had never used illegal drugs, but later in the questionnaire, when asked how many times they had used marijuana, they answered that question indicating a number greater than 0. Thus, what should have been a question that was legitimately skipped was answered with an unexpected number. What could this mean? One possibility is that the respondent was not paying attention to the questions and answered carelessly or in error. Another possibility is that the initial answer (have you ever used illegal drugs) was answered incorrectly. It also is possible that some subset of the population did not include marijuana in the category of illegal drugs—an interesting finding in itself and one way in which researchers can use data cleaning to improve their subsequent research.

Legitimate missing data can be dealt with in different ways. One common way of dealing with this sort of data could be using analyses that do not require (or can deal effectively with) incomplete data. These include things like hierarchical linear modeling (HLM) (Raudenbush & Bryk, 2002) or survival analysis.[2] Another common way of dealing with this sort of legitimate missing data is adjusting the denominator (an important concept introduced in Chapter 3). Again taking the example of the marriage survey, we could eliminate nonmarried individuals from the particular analysis looking at length of marriage, but would leave nonmarried respondents in the analysis when looking at issues relating to being married versus not being married. Thus, instead of asking a slightly silly question of the data—"How long, on average, do all people, even unmarried people, stay married?"—we can ask two more refined questions: "What are the predictors of whether someone is currently married?" and "Of those who are currently married, how long on average have they been **[p. 108 ↓]** married?" In this case, it makes no sense to include nonmarried individuals in the data on how long someone has been married.

This example of dealing with legitimately missing data is relatively straightforward and mostly follows common sense. The best practice here is to make certain the denominator (the sample or subsample) is appropriate for the analysis. Be sure to report having selected certain parts of your sample for specific analyses when doing so. In the case of legitimate missing data, it is probably rare that a researcher would want to deal with it by imputing or substituting a value (as we discuss for illegitimately missing data below), as that again changes the research question being addressed to "If everyone was married, how long, on average, would they stay married?" That probably is not something that makes a tremendous amount of sense.

Illegitimately missing data is also common in all types of research. Sensors fail or become miscalibrated, leaving researchers without data until that sensor is replaced or recalibrated. Research participants choose to skip questions on surveys that the researchers expect everyone to answer. Participants drop out of studies before they are complete. Missing data also, somewhat ironically, can be caused by data cleaning. It is primarily this second type of missing data that I am most concerned with, as it has the potential to bias the results.

Few authors seem to explicitly deal with the issue of missing data, despite its obvious potential to substantially skew the results (Cole, 2008). For example, in a recent survey my students and I performed of highly regarded journals from the American

Psychological Association, we found that more than one-third (38.89%) of authors discussed the issue of missing data in their articles (Osborne, Kocher, & Tillman, 2011). Do those 61% who fail to report anything relating to missing data have complete data (rare in the social sciences, but possible for some authors), do they have complete data because they removed all subjects with any missing data (undesirable, and potentially biasing the results, as we discuss next), did they deal effectively with the missing data and fail to report it (less likely, but possible), or did they allow the statistical software to treat the missing data via whatever the default method is, which most often leads to deletion of subjects with missing data? If our survey is representative of researchers across the sciences, we have cause for concern. Our survey found that of those researchers who did report something to do with missing data, most reported having used the classic methods of listwise deletion (complete case analysis) or mean substitution, neither of which are **[p. 109 ↓]** particularly effective practices (Schafer & Graham, 2002), as I demonstrate below. In only a few cases did researchers report doing anything constructive with the missing data, such as estimation or imputation. And in no case did we find that researchers analyzed the missingness to determine whether it was *missing completely at random* (MCAR), *missing at random* (MAR), or *missing not at random* (MNAR). This suggests there is a mythology in quantitative research that (a) individuals with incomplete data cannot contribute to the analyses, and that (b) removing them from the analyses is an innocuous action, which is only justified if you believe that missing data is missing completely at random (probably not the most common state).

Categories of Missingness

When exploring missing data, it is important to come to a conclusion about the *mechanism of missingness*—that is, the hypothesized reason for why data are missing. This can range from arbitrary or random influences to purposeful patterns of nonresponse (e.g., most women in a study refuse to answer a question that is offensive or sensitive to women but that does not affect men in the same way).

Determination of the mechanism is important. If we can infer the data are missing at random (i.e., MCAR or MAR), then the nonresponse is deemed *ignorable*. In other words, random missingness can be problematic from a power perspective (in that it often reduces sample size or degrees of freedom for an analysis), but it would not potentially bias the results. However, data missing *not* at random (MNAR) could potentially be a strong biasing influence (Rubin, 1976).

Let us take an example of an employee satisfaction survey given to schoolteachers in a local district as an example of MCAR, MAR, and MNAR. Imagine that in September all teachers are surveyed (X), and then in January teachers are surveyed again (Y). Missing completely at random (MCAR) would mean that missingness in January is completely unrelated to any variable, including September satisfaction level, age, years of teaching, and the like. An example of this would be 50% of all respondents from September were randomly sampled to respond to the survey again in January, with all potential respondents completing surveys at both time points. In this case, having data for Y present or absent is completely explained by random selection. Put **[p. 110 ↓]** another way, missingness has no systematic relation to any variable present or unmeasured (such as age, sex, race, level of satisfaction, years teaching).

Now imagine that this surveying was part of the school district's initiative to keep teachers from leaving, and they wanted to focus on teachers with low satisfaction in September, perhaps with an intervention to help raise satisfaction of these low-satisfaction teachers. In this case, the missingness depends solely and completely on X, the initial score. Because the goal of the survey is to explore how these particular teachers fared, rather than all teachers in general, missingness is still considered ignorable and missing at random (MAR). If, on the other hand, other factors aside from initial satisfaction level were responsible (or partly responsible for missingness) such that perhaps only teachers whose satisfaction had improved responded (the teachers who continued to be substantially dissatisfied may be less likely to return the survey), then the data are considered missing not at random (MNAR) and are not ignorable (Rubin, 1976; Schafer & Graham, 2002) because they may substantially bias the results. In the case of MNAR, the average satisfaction of the follow-up group would be expected to be inflated if those who were most dissatisfied had stopped responding. If missingness were related to another external factor, such as if those teachers who were most dissatisfied were the most junior teachers (the teachers with least time in the profession), that also would qualify the missing data as MNAR.

In other words, it is only legitimate to assume that your observed data are representative of the intended population if data are convincingly missing at random or missing completely at random.[3] For simplicity, I will proceed through the rest of the chapter focusing on MCAR versus MNAR. MAR (ignorable missingness) is probably more common than MCAR but MNAR is probably most common, and thus, MCAR is merely presented as a comparison point. In truth, best practices in handling missing data appear to be equally effective regardless of whether the data are MCAR, MAR, or MNAR.

What do we do with Missing Data?

To illustrate some of the effects of missing data handling, I used data from the Education Longitudinal Study of 2002 (Ingels et al., 2004), grade 10 cohort to provide an example. For these analyses, no weights were applied. The complete sample of 15,163 students represents our example of the population (the advantage here is that we know the exact parameters of the population, **[p. 111 ↓]** something we often do not know). In this first example, I use the relatively strong correlation between math and reading achievement scores (BYTXMIRR, BYTXRIRR), which produces what we define as the "population" correlation estimate (ρ) of .77, as indicated in Table 6.1 (row #1). (See also Figure 6.3 on page 134.)

Data Missing Completely at Random (MCAR)

To simulate MCAR situations, 20% of mathematics scores were randomly selected to be identified as missing. As a confirmation of the randomness of the missingness, two analyses were performed. First, as Table 6.1 shows, there was no mean difference in reading IRT scores between the missing and non-missing groups (*F*
(1, 1516)

= 0.56, $p < .45$, $\eta^2 = .0001$). Second, there was no correlation between the missingness variable and any other substantive or ancillary variable (e.g., socioeconomic status, standardized reading IRT scores; all *r*
(15,163)

.002 to .006, $p < .57$ to .79). Another test of randomness was a logistic regression predicting missingness (0 = not missing, 1 = missing) from all other variables (math, reading, and socioeconomic status). When all three variables were in the equation, the overall equation was not significant ($p < .47$) and all 95% confidence intervals for the odds ratios for the three variables included 1.00, indicating no significant relationship between missingness and any of the three variables.[4] Finally, another test of randomness is to perform an ANOVA to see if individuals with missing data on one variable are significantly different on other, similar variables (in this case, reading achievement). As you can see in Table 6.1, there is no significant difference in reading achievement between those with missing data on math achievement and those with valid math scores. Although not definitive, this sort of analysis in your data can give support to an inference of randomness or nonrandomness regarding the missing data.

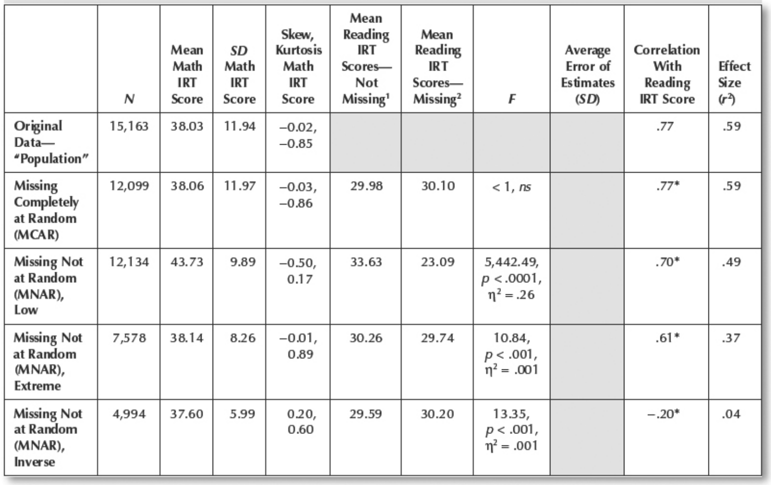
Data Missing Not at Random—Low Scoring Students More Likely to Be Missing (MNAR-Low)

To simulate one type of MNAR (labeled MNAR-low), cases at or below the 30th percentile on the math achievement test were given a 80% chance of being randomly labeled as missing on the math test, cases between the 30th and 50th percentile on the math test were given a 50% chance of being

[p. 112 ↓]

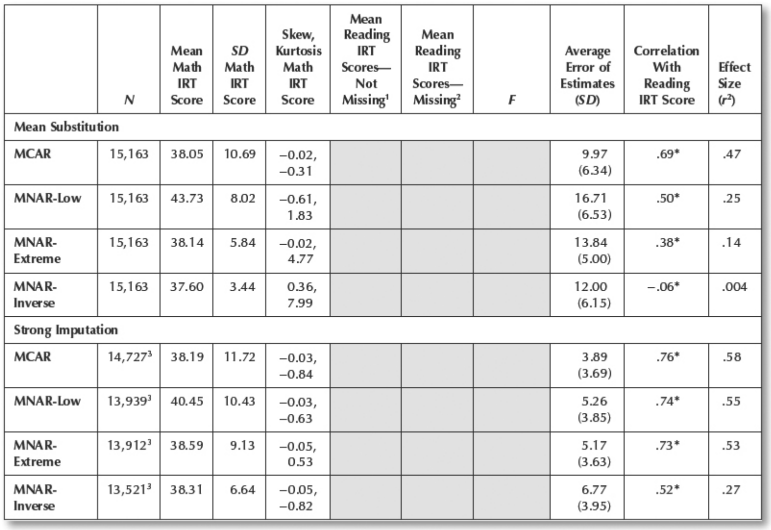
Table 6.1 Summary of Effects of Missingness Corrections for Math Achievement Scores

Table 6.1 Summary of Effects of Missingness Corrections for Math Achievement Scores



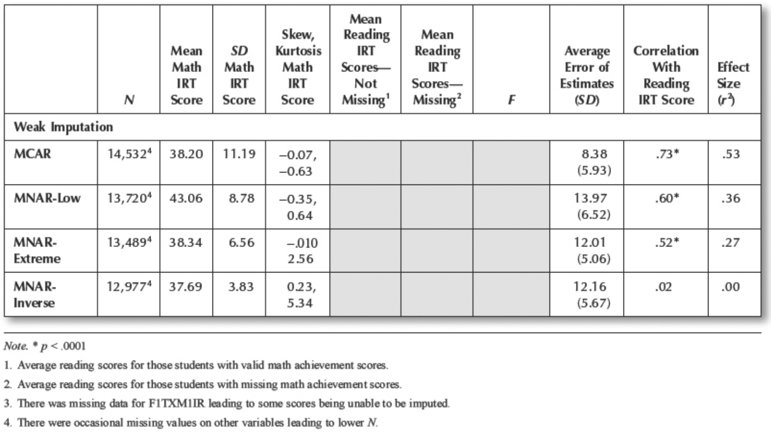
	N	Mean Math IRT Score	*SD* Math IRT Score	Skew, Kurtosis Math IRT Score	Mean Reading IRT Scores—Not Missing[1]	Mean Reading IRT Scores—Missing[2]	*F*	Average Error of Estimates (*SD*)	Correlation With Reading IRT Score	Effect Size (r^2)
Original Data—"Population"	15,163	38.03	11.94	−0.02, −0.85					.77	.59
Missing Completely at Random (MCAR)	12,099	38.06	11.97	−0.03, −0.86	29.98	30.10	< 1, *ns*		.77*	.59
Missing Not at Random (MNAR), Low	12,134	43.73	9.89	−0.50, 0.17	33.63	23.09	5,442.49, $p < .0001$, $\eta^2 = .26$		.70*	.49
Missing Not at Random (MNAR), Extreme	7,578	38.14	8.26	−0.01, 0.89	30.26	29.74	10.84, $p < .001$, $\eta^2 = .001$		.61*	.37
Missing Not at Random (MNAR), Inverse	4,994	37.60	5.99	0.20, 0.60	29.59	30.20	13.35, $p < .001$, $\eta^2 = .001$		−.20*	.04

[p. 113 ↓]



	N	Mean Math IRT Score	*SD* Math IRT Score	Skew, Kurtosis Math IRT Score	Mean Reading IRT Scores—Not Missing[1]	Mean Reading IRT Scores—Missing[2]	*F*	Average Error of Estimates (*SD*)	Correlation With Reading IRT Score	Effect Size (r^2)
Mean Substitution										
MCAR	15,163	38.05	10.69	−0.02, −0.31				9.97 (6.34)	.69*	.47
MNAR-Low	15,163	43.73	8.02	−0.61, 1.83				16.71 (6.53)	.50*	.25
MNAR-Extreme	15,163	38.14	5.84	−0.02, 4.77				13.84 (5.00)	.38*	.14
MNAR-Inverse	15,163	37.60	3.44	0.36, 7.99				12.00 (6.15)	−.06*	.004
Strong Imputation										
MCAR	14,727[3]	38.19	11.72	−0.03, −0.84				3.89 (3.69)	.76*	.58
MNAR-Low	13,939[3]	40.45	10.43	−0.03, −0.63				5.26 (3.85)	.74*	.55
MNAR-Extreme	13,912[3]	38.59	9.13	−0.05, 0.53				5.17 (3.63)	.73*	.53
MNAR-Inverse	13,521[3]	38.31	6.64	−0.05, −0.82				6.77 (3.95)	.52*	.27

[p. 114 ↓]



	N	Mean Math IRT Score	*SD* Math IRT Score	Skew, Kurtosis Math IRT Score	Mean Reading IRT Scores—Not Missing[1]	Mean Reading IRT Scores—Missing[2]	*F*	Average Error of Estimates (*SD*)	Correlation With Reading IRT Score	Effect Size (r^2)
Weak Imputation										
MCAR	14,532[4]	38.20	11.19	−0.07, −0.63				8.38 (5.93)	.73*	.53
MNAR-Low	13,720[4]	43.06	8.78	−0.35, 0.64				13.97 (6.52)	.60*	.36
MNAR-Extreme	13,489[4]	38.34	6.56	−.010 2.56				12.01 (5.06)	.52*	.27
MNAR-Inverse	12,977[4]	37.69	3.83	0.23, 5.34				12.16 (5.67)	.02	.00

Note. * $p < .0001$

1. Average reading scores for those students with valid math achievement scores.
2. Average reading scores for those students with missing math achievement scores.
3. There was missing data for F1TXM1IR leading to some scores being unable to be imputed.
4. There were occasional missing values on other variables leading to lower *N*.

[p. 115 ↓]

randomly labeled as missing, and those over the 50th percentile on the math test were only given a 1% chance of being labeled as missing on the math test. This should simulate a highly biased situation where the best-performing students are more likely to respond to an item than the worst-performing students. As expected, MNAR-low produced an upwardly biased estimate of average performance—overestimating the mean performance due to more missing data from lower-performing students and slightly underestimating the standard deviation, also expected in this case due to less dispersion at the lower extreme of the distribution. As expected, achievement test scores were significantly correlated with missingness in this case (*r*
(15,163)

= -.51 and -.66 for reading and math achievement, respectively, both *p* < .0001) as was socioeconomic status (*r*
(15,163)

= -.28, *p* < .0001). Furthermore, logistic regression predicting missingness from achievement and socioeconomic status found all three variables were significant predictors of MNAR-low (all *p* < .0001), indicating that those with lower achievement (or SES) were more likely to be missing (as expected). Finally, as Table 6.1 shows, there were substantial mean differences in reading achievement between those with missing scores and those with valid math scores.

Data Missing Not at Random—Students at the Extremes More Likely to Be Missing (MNAR-Extreme)

A second type of MNAR (MNAR-extreme) was simulated by giving those students below the 30th percentile and above the 70th percentile on the math achievement test an 80% chance of being randomly identified as missing on the math test. Those in the center of the distribution were given only a 5% chance of being labeled as missing on the math test (Acock, 2005). This should have the effect of increased nonrandom missingness without substantially skewing the population average estimates.

As expected, MNAR-extreme produced the desired effects. Because the highest and lowest 30% of the students were more likely to be missing than the middle 40% (i.e., the missing data was symmetrically, but not randomly distributed), the distribution should closely match the mean of the original population, with dramatically reduced variance, and little or no difference in missing or nonmissing scores. As Table 6.1 shows, that is exactly what occurred. The average for MNAR-extreme closely approximates the population mean, underestimates the standard deviation, and produced significant, but unimportant differences between the two groups (an eta-squared of .001 is an extremely small effect size). Furthermore, we would not expect significant correlations **[p. 116 ↓]** between missingness and achievement or socioeconomic status, and correlations ranged from *r*
(15,163)

= -.03 to .02. Finally, though the logistic regression indicated that missingness in this case was significantly related to reading achievement (Odds ratio = 0.99, $p < .0001$) and socioeconomic status (Odds ratio = 1.10, $p < .0001$), the odds ratios are close to 1.00, indicating a small effect size that is only significant by virtue of having more than 15,000 degrees of freedom. Thus, I can argue that while MNAR-extreme was decidedly non-random missingness, it did produce a largely symmetrical distribution.

Complete case analysis can lead to incomplete understanding.

Stuart, Azur, Frangakis, and Leaf (2009) give some interesting examples of how looking at only cases with complete data can lead to incomplete or inaccurate findings in the context of a national health survey. In one example, eliminating cases with missing data could lead us to conclude that individuals who start smoking earlier in life are more emotionally strong and less functionally impaired than individuals who started smoking later in life—a finding contrary to common sense and decades of research. They also found that under complete case analysis, those who drink more have *fewer internalizing problems* (e.g., depression, anxiety), another incongruous finding. Fortunately, after appropriate handling of missing data, these relationships were more consistent with the literature.

These real-life examples inspired me to create the fourth condition, MNAR-inverse because missing data apparently can lead to completely wrong conclusions in the real world.

Data Missing not at Random that Inverts the Relationship between Two Variables (MNAR-Inverse)

As a final challenge and test of missing data handling techniques, I created an extremely biased sampling technique that virtually eliminated those with both high reading and math scores, and those with both low reading and math scores, to have the effect of reversing the relationship between reading and math achievement (this is described more thoroughly in Appendix A of this chapter and also is available on the book's website). (See also Figure 6.4 on page 135.) By selectively sampling only those students on the downward diagonal, this produced a sample of almost N = 5,000 students that had a negative correlation (r

(4,994)

= -.20).

Finally, MNAR-inverse also had the desired effect of producing a sample that at a glance does not look problematic. As Table 6.1 (5th row) shows, this MNAR-inverse sample is not substantially different from the other samples in mean math achievement (although the standard deviation **[p. 117 ↓]** underestimates the population variablility), and the shape of the distribution is not substantially different from the MNAR-extreme distribution. Furthermore, there is little difference between those missing and not missing on the reading achievement score (again, a very small effect size of eta-squared = .001). Other analyses showed no important correlations between missingness and achievement or socioeconomic status (*r*
(15,163)

ranged from .03 to .04), and a logistic regression predicting missingness from the same three variables showed only a small effect for socioeconomic status (Odds ratio = 1.08, *p* < .0001) indicating that those from more affluent families were more likely to be missing. If a researcher was unaware that the population correlation for these two variables should be .77, none of these minor effects hint at how biased this sample is due to nonrandom missingness—yet this example highlights the importance of dealing effectively with missing data.

The Effects of Listwise Deletion

Traditional methods of dealing with missing data (and the default for many statistical packages) is to merely delete any cases with missing values on any variable in the analysis. A special case of this, called *pairwise deletion* or *available case analysis*, uses those cases with complete data on only those variables selected for a particular analysis. This means that the sample being analyzed can change depending on which variables are in the analysis, which could be problematic regarding replicability and increase the odds of errors of inference. Neither case is particularly desirable (Cole, 2008; Schafer & Graham, 2002). When data are MCAR, estimates are not biased, but under the more common MAR or MNAR conditions, misestimation and errors can result

(Stuart, et al., 2009). Again, referring to Table 6.1, a simple example of the correlation between reading and math achievement test scores demonstrates this effect nicely.

As Table 6.1 shows, the original correlation coefficient for the population was ρ = .77 (variance accounted for = .59). When the data are MCAR, the population effect is estimated almost exactly. However, when data are MNAR, estimates begin to stray from the population parameter. In the MNAR-low sample, what might look like a minor misestimation (*r*
(12,134)

= .70) is an underestimation of the effect size by almost 20% (coefficients of determination/percentage variance accounted for are .59 versus .49, a 16.9% underestimation). When the missing data causes a restriction of range situation (introduced in Chapter 3, showing restriction **[p. 118 ↓]** of range causing attenuation of correlation coefficients) represented by the MNAR-extreme sample, the misestimation is even more pronounced, producing a correlation coefficient of *r*
(7,578)

= .61 (coefficient of determination of 0.37, which underestimates the population effect size by 37.29%). Finally, and most obviously, when the missingness is biased in a particular way, such as the MNAR-inverse example, it is possible that deletion of cases could lead researchers to draw the opposite conclusion regarding the nature of the relationship than exists in the population, as evidenced by the MNAR-inverse sample.

Thus, by deleting those with missing data, a researcher could be misestimating the population parameters, making replication less likely (for more examples of this effect, see Schafer & Graham, 2002, Table 2.).

Another undesirable effect of case deletion (even under MCAR) is loss of power. Most researchers use analyses with multiple variables. If each variable has some small percentage of randomly missing data, five variables with small percentages of missing data can add up to a substantial portion of a sample being deleted, which can have deleterious effects on power (as discussed in Chapter 2). Combined with what is likely an underestimation of the effect size, power can be significantly impacted when substantial portions of the sample are deleted when data are not MCAR. Thus, case deletion is only an innocuous practice when (a) the number of cases with missing data is a small percentage of the overall sample, and (b) the data are *demonstrably* MAR.

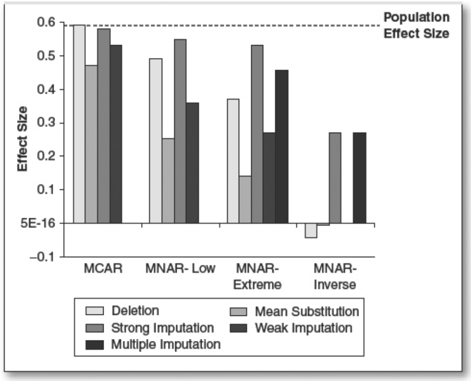
The Detrimental Effects of Mean Substitution

I have seen two types of mean substitution. In one case, an observed variable (e.g., number of years of marriage) is unreported, the group or overall sample mean is substituted for each individual with missing data. The theory is that, in the absence of any other information, the mean is the best single estimate of any participant's score. The flaw in this theory is that if 20% of a sample is missing, even at random, substituting the identical score for a large portion of the sample artificially reduces the variance of the variable, and as the percentage of missing data increases, the effects of missing data become more profound. These effects have been known for many decades now (Cole, 2008; Haitovsky, 1968), yet many researchers still view mean substitution as a viable, or even progressive, method of dealing with missing data. As you will see below (and in Figure 6.1), mean substitution can create more inaccurate population estimates than simple case deletion when data are not MCAR.

[p. 119 ↓]

To simulate this effect as a real researcher would face it, I substituted the mean of the math achievement variable calculated once the missing values were inserted into the variable.[5] As Table 6.1 shows, standard deviations are underestimated under MNAR situations.[6] For example, even under MCAR, the variability of math achievement is underestimated by 10.47% when mean substitution is used (and the effect would become more substantial as a larger percentage of the sample were missing), although the estimate of the mean is still accurate. In this case, the correlation effect size also is underestimated by 20.34% (coefficient of determination = 0.59 versus 0.47) just through virtue of 20% of the sample being MCAR and substituting the mean to compensate. Note also that mean substitution under MCAR appears to be less desirable than case deletion. In Figure 6.1, comparing MCAR with deletion and MCAR with mean substitution, you can see that the estimates of the population are more accurate when the missing cases are deleted.

Figure 6.1 Under Estimation of Effect Size When Missing Data are Treated Via Deletion, Mean Substitution, Strong Imputation, Weak Imputation, and Multiple Imputation



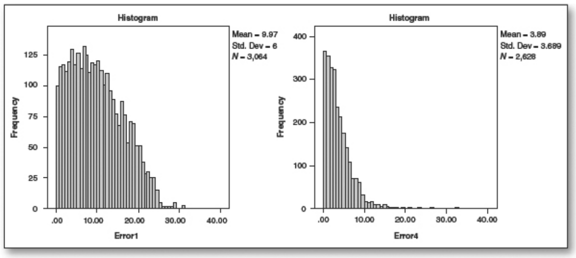
[p. 120 ↓]

To explore the errors mean substitution made, even under MCAR, the difference between the mean substituted and the original score was calculated and is presented in Figure 6.2. As you might expect from randomly missing data, the average error is almost 0 (-0.17), but there is a large range (-25.54 to 31. 15). Taking the absolute values of each error (presented in Figure 6.2), the average error of estimating scores via mean substitution is 9.97 with a standard deviation of 6.34.

The effects of mean substitution appear more dramatic under MNAR-low, despite being approximately the same overall number of missing cases. This is because the missing data in this case are likely to be low-performing students, and the mean is a poor estimate of their performance (average error in this case is 16.71, standard deviation is 6.53, much larger than under MCAR). Thus, under MNAR-low, mean substitution produces a biased mean, substantially underestimates the standard deviation by almost 33%, dramatically changes the shape of the distribution (skew, kurtosis), and leads to significant underestimation of the correlation between reading and math achievement. Under MNAR-low with mean substitution, the effect size for this simple correlation is underestimated by 57.63% (coefficient of determination = 0.59 versus 0.25). Note that MNAR with deletion produced better population estimates than mean substitution.

The example of MNAR-extreme also exposes the flaws of mean substitution. Note that because the missing data were symmetrical, the estimate of the **[p. 121 ↓]** population mean was excellent both when cases were deleted and when mean substitution was used. However, the case of MNAR-extreme with mean substitution produced inaccurate estimates of population variability (*SD* = 5.84 versus population *SD* = 11.94, a 51.09% underestimation of the true variability in the population). Again, this is not a surprise as the average error from mean substitution is 13.84, standard deviation is 5.00. Furthermore, because of the high concentration of missing data in the tails of the distribution, the shape of the distribution becomes dramatically nonnormal. Finally, the effect size of the simple correlation between reading and math achievement scores is underestimated by 76.27% (0.59 versus 0.14), a notably poorer estimation than merely deleting cases under MNAR-extreme.

Figure 6.2 Misestimation of Math Scores Under Mean Substitution, Strong Imputation, MCAR



It should be no surprise that mean substitution does little to help the situation of MNAR-inverse. The correlation is simply a complete misestimation of the population parameter, has high error (but not as high as the two other MNAR samples, interestingly), and substantially underestimates population variability. Thus, this type of mean substitution does not appear to be an acceptable practice in which researchers should engage.

Mean substitution when creating composite scores based on multi-item questionniares. The other type of mean substitution involves administration of psychological scales (e.g., self-esteem, depression) where there are multiple, highly correlated questions assessing a single construct. In the case of the Rosenberg SVI, for example, where internal reliability estimates are often in excess of .90, the theory is that it is more desirable to substitute that individual's mean for the other items rather than to discard the individual from the data set. Thus, the idea that significant information is contained in the other highly correlated answers is an intriguing one, and used to generate other estimates discussed below. In this case, as item intercorrelations get higher, and the number of items increases, the bias does not appear to be substantial (Schafer & Graham, 2002), but this holds true only if the scale is unidimensional. In other words, if a scale has multiple independent aspects or subscales (e.g., depression is often not considered a unitary scale, and therefore averaging all the items would not be appropriate) it is only legitimate to average the items from the subscale the missing value belongs to.[7] This type of mean substitution is similar to imputation, discussed next, and when the imputation is based on strong relationships, it can be very effective. Thus, this type of mean substitution for missing scale items when internal consistency is strong and the scale **[p. 122 ↓]** is unidimensional appears to be a defensible practice. Of course, measurement scholars will argue that there are more modern methods of dealing with this sort of issue, and they are correct. If you are trained in more advanced measurement techniques, please use them.

The Effects of Strong and Weak Imputation of Values

Conceptually, the second type of mean substitution mentioned earlier is similar to imputation via multiple regression. It uses information available in the existing data to estimate a better value than the sample average, which as we saw in the previous section, is only effective at reducing the accuracy of the analysis. Essentially, imputation combines the complexity of predictive applications of multiple regression, which I think is excellently discussed in an article I wrote and which is freely available on the Internet (Osborne, 2000). In practice, assuming most variables have complete data for most participants, and they are strongly correlated to the variable with the missing data, a researcher can create a prediction equation using the variables with complete data, estimating values for the missing cases much more accurately than simple mean substitution.

To demonstrate this under the most ideal circumstances, I used two variables from the ELS 2002 data set that are correlated with the 10th grade math achievement variable that contains missing values: 12th grade math achievement (F1TXM1IR) and socioeconomic status (BYSES2). As imputation involves creating a regression equation based on the valid cases in a sample, for each simulation below I used only cases with nonmissing data to generate the regression equation, as a researcher faced with a real data set with missing data would have to do. For reference, I also calculated the regression equation for the population. These equations represent strong imputation, as the variance accounted for is very high (.40 to .80).

Prediction equation for population:

$$\text{Math} = 5.286 + 0.552(\text{BYSES2}) + 0.680(\text{F1TXM1IR})\ (r^2 = .80).$$

Prediction equation for MCAR sample:

$$\text{Math} = 5.283 + 0.533(\text{BYSES2}) + 0.681(\text{F1TXM1IR})\ (r^2 = .80).$$

[p. 123 ↓]

Prediction equation for MNAR-low sample:

$$\text{Math} = 9.907 + 0.437(\text{BYSES2}) + 0.617(\text{F1TXM1IR})\ (r^2 = .72).$$

Prediction equation for MNAR-extreme sample:

$$\text{Math} = 11.64 + 0.361(\text{BYSES2}) + 0.548(\text{F1TXM1IR})\ (r^2 = .63).$$

Prediction equation for MNAR-inverse sample:

$$\text{Math} = 18.224 + -0.205(\text{BYSES2}) + 0.407(\text{F1TXM1IR})\ (r^2 = .40).$$

It should not be surprising that the prediction equations became increasingly less similar to the population equation (and less effective) as I moved from MCAR to MNAR-low to MNAR-extreme to MNAR-inverse. However, given the extremely high predictive power of 12th grade math achievement scores in predicting 10th grade math achievement (*r (12,785)*

= .89, which has a coefficient of determination of 0.79), prediction even in the worst case is strong. The relevant question is whether these equations will produce better estimations than mean substitution or complete case analysis.

As Table 6.1 and Figure 6.1 show, given this strong prediction, under MCAR the population mean and standard deviation, as well as the distributional properties, are closely replicated. Under MNAR-low, MNAR-extreme, and MNAR-inverse, the misestimation is significantly reduced, and the population parameters and distributional properties are more closely approximated than under mean substitution. Further, in all cases the errors of the estimates dropped markedly (as one might expect using such powerful prediction rather than mean substitution). Finally, under imputation, the estimates of the correlation between reading and math achievement test scores are much closer to approximating the population correlation than either deletion or mean substitution. This is particularly true for MNAR-inverse, where we see the true power of more progressive missing value handling techniques. Researchers using strong imputation would estimate a relationship between these two variables in the correct direction and, while underestimated, it is much closer to the population parameter than under any other technique.

Unfortunately, it is not always the case that one has another variable with a correlation of this magnitude with which to predict scores for missing values. Thus, to simulate a weaker prediction scenario, I used other variables from the same data set: BYSES2 (socioeconomic status), BYRISKFC (number of **[p. 124 ↓]** academic risk factors a student exhibits), and F1SEX (1 = male, 2 = female). Collectively, these three variables represent modest predictive power, with an $r = .49$, $r^2 = .24$, $p < .0001$ for the model. The predictive equations are as follows: Prediction equation for population:

$$\text{Math} = 42.564 + 5.487(\text{BYSES2}) - 1.229(\text{F1SEX}) - 2.304(\text{BYRISKFC})\ (r^2 = .24).$$

Prediction equation for MCAR sample:

$$\text{Math} = 42.701 + 5.468(\text{BYSES2}) - 1.241(\text{F1SEX}) - 2.368(\text{BYRISKFC})\ (r^2 = .24).$$

Prediction equation for MNAR_low sample:

$$\text{Math} = 46.858 + 4.035(\text{BYSES2}) - 1.440(\text{F1SEX}) - 1.748(\text{BYRISKFC})\ (r^2 = .17).$$

Prediction equation for MNAR_extreme sample:

$$\text{Math} = 40.149 + 3.051\ (\text{BYSES2}) - 0.491(\text{F1SEX}) - 1.155(\text{BYRISKFC})\ (r^2 = .13).$$

Prediction equation for MNAR_inverse sample:

$$\text{Math} = 40.416 + 0.548\ (\text{BYSES2}) - 1.460(\text{F1SEX}) - 0.547(\text{BYRISKFC})\ (r^2 = .03).$$

As you can see from this more realistic example (Table 6.1, and Figure 6.1), as the imputation gets weaker, the results get closer to mean substitution. In this case, the prediction was generally better than simple mean substitution, but not as good as strong imputation. As Table 6.1 shows, under MNAR-low, MNAR-extreme, and MNAR-inverse conditions, the variance of the population was misestimated, and in the case of MNAR-low, the population mean also was misestimated. The errors of estimation, while not as large as mean substitution, were still undesirably large. Finally, estimation of the population correlation between math and reading achievement tests were improved over mean substitution, but still misestimated compared to strong imputation.

So where does that leave us? Under the best circumstances, imputation appears to give the best results, even correcting the undesirable situation present **[p. 125 ↓]** in MNAR-inverse, particularly when prediction is strong and done well. When done poorly, imputation can cause distortion of estimates and lead to errors of inference (Little & Rubin, 1987), just as complete case analysis can (Stuart et al., 2009). In large samples with strongly correlated variables and low rates of missing data, this appears to be a good option from amongst the classic techniques thus far, although far more effective when data are missing at random than when missingness is biased in some way. However, recent research has shown that taking the extra effort of using advanced, modern estimation procedures can have benefits for those researchers with relatively high rates of missingness. It is beyond the scope of this chapter to get into all the details of all these different advanced techniques, but I will briefly address one of the more common ones for those curious in exploring further.

Multiple Imputation: A Modern Method of Missing Data Estimation

Multiple imputation (MI) has emerged as one of the more common modern options in missing data handling with the ubiquity of desktop computing power. Essentially, multiple imputation uses a variety of advanced techniques—e.g., EM/maximum likelihood estimation, propensity score estimation, or Markov Chain Monte Carlo (MCMC) simulation—to estimate missing values, creating multiple versions of the same data set (sort of a statistician's view of the classic science fiction scenario of alternate realities or parallel universes) that explore the scope and effect of the missing data. These parallel data sets can then be analyzed via standard methods and results combined to produce estimates and confidence intervals that are often more robust than simple (especially relatively weak) imputation or previously mentioned methods of dealing with missing values (Schafer, 1997, 1999).

When the proportion of missing data is small and prediction is good, single imputation described above is probably sufficient, although as with any prediction through multiple regression, it “overfits” the data, leading to less generalizable results than the original data would have (Osborne, 2000, 2008; Schafer, 1999).[8] The advantage of MI is generalizability and replicability—it explicitly models the missingness and gives the researcher confidence intervals for estimates rather than trusting to a single imputation. Some statistical software packages are beginning to support MI (e.g., SAS, R, S-Plus, SPSS—with **[p. 126 ↓]** additionally purchased modules and standalone software such as that available from Joseph Schafer at http://www.stat.psu.edu/~jls/software.html). Finally, and importantly, some MI procedures do *not* require that data be missing at random (e.g., in SAS there are several options for estimating values depending on the assumptions around the missing data). In other words, under a worst-case scenario of a substantial portion of missing data that is due to some significant bias, this procedure should be a good alternative (Schafer, 1999).

I used SAS's PROC MI procedure as it is relatively simple to use (if you are at all familiar with SAS)[9] and has the nice option of automatically combining the multiple parallel data sets into one analysis. For this analysis, I prepared a data set that contained the math and reading achievement test scores, as well as the three variables used for weak imputation (sex of student, socioeconomic status, and risk factors), and used the SAS defaults of EM estimation with five parallel data sets.

The traditional view within multiple imputation literature has been that five parallel data sets is generally a good number, even with high proportions of missing data. More recent studies suggest that 20 should be a minimum number of iterations (Graham, Olchowski, & Gilreath, 2007). In truth, with software that can perform MI automatically, there is no reason *not* to do more iterations. But in the case of this analysis, five parallel data sets achieved a relative efficiency of 96%, a good indicator. For illustrative purposes, Table 6.2 shows the five different imputations.

As you can see in Table 6.2 (and Figure 6.1), even using the weak relationships between the variables from the weak imputation example, the results are much better than the simple weak imputation (closer to strong imputation) and remarkably consistent. And the variance of the population, the shape of the variable distribution, and the estimation of the correlation between the two variables of interest are estimated much more accurately than any other method save having an extremely highly correlated variable to help with imputation. These estimates would then be combined to create a single estimate of the effect and confidence interval around that effect. In this case, the effect was so consistent that step was not necessary for this purpose.

Can multiple imputation fix the highly biased missingness in MNAR-inverse? As a final test of the power of MI, Table 6.3 shows the 20 EM imputations I performed to get a relative efficiency of 98% on the MNAR-inverse data.[10] By analyzing the 20 imputations through PROC MI ANALYZE, SAS provides the

[p. 127 ↓]

Table 6.2 Example of Multiple Imputation Using Sas Proc MI and Weak Predictors Only, MNAR-Extreme Missingness Pattern

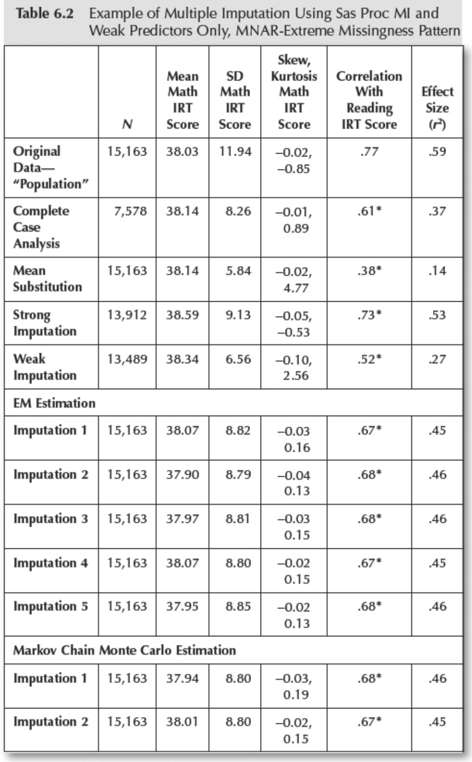
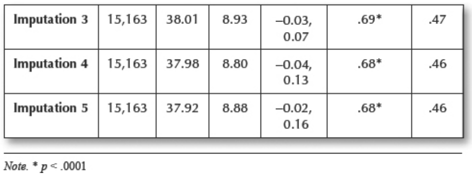


Table 6.2 Example of Multiple Imputation Using Sas Proc MI and Weak Predictors Only, MNAR-Extreme Missingness Pattern

	N	Mean Math IRT Score	SD Math IRT Score	Skew, Kurtosis Math IRT Score	Correlation With Reading IRT Score	Effect Size (r^2)
Original Data—"Population"	15,163	38.03	11.94	–0.02, –0.85	.77	.59
Complete Case Analysis	7,578	38.14	8.26	–0.01, 0.89	.61*	.37
Mean Substitution	15,163	38.14	5.84	–0.02, 4.77	.38*	.14
Strong Imputation	13,912	38.59	9.13	–0.05, –0.53	.73*	.53
Weak Imputation	13,489	38.34	6.56	–0.10, 2.56	.52*	.27
EM Estimation						
Imputation 1	15,163	38.07	8.82	–0.03 0.16	.67*	.45
Imputation 2	15,163	37.90	8.79	–0.04 0.13	.68*	.46
Imputation 3	15,163	37.97	8.81	–0.03 0.15	.68*	.46
Imputation 4	15,163	38.07	8.80	–0.02 0.15	.67*	.45
Imputation 5	15,163	37.95	8.85	–0.02 0.13	.68*	.46
Markov Chain Monte Carlo Estimation						
Imputation 1	15,163	37.94	8.80	–0.03, 0.19	.68*	.46
Imputation 2	15,163	38.01	8.80	–0.02, 0.15	.67*	.45

[p. 128 ↓]



Imputation 3	15,163	38.01	8.93	–0.03, 0.07	.69*	.47
Imputation 4	15,163	37.98	8.80	–0.04, 0.13	.68*	.46
Imputation 5	15,163	37.92	8.88	–0.02, 0.16	.68*	.46

Note. * $p < .0001$

average of the estimate, the standard error of the estimate, 95% confidence interval for the estimate, and more. In this case, the 20 iterations produced an average standardized regression coefficient (identical to correlation in this example) of 0.51, with a standard error of 0.00982, a 95% confidence interval of 0.49 to 0.52.

Ultimately, multiple imputation (and other modern missing value estimation techniques) are increasingly accessible to average statisticians and therefore represents an exciting frontier for improving data cleaning practice. As the results in Tables 6.2 and 6.3 show, even with only modestly correlated variables and extensive missing data rates, the MI techniques demonstrated here gave superior results to single, weak imputation for the MNAR-extreme and MNAR-inverse missingness patterns. These represent extremely challenging missingness issues often not faced by average researchers, but it should be comforting to know that appropriately handling missing data, even in extremely unfortunate cases, can still produce desirable (i.e., accurate, reproducible) outcomes. MI techniques seem, therefore, to be vastly superior to any other, traditional technique. Unfortunately, no technique can completely recapture the population parameters when there are such high rates of missing-ness, and in such a dramatically biased fashion. But these techniques would at least keep you, as a researcher, on safe ground concerning the goodness of inferences you would draw from the results.

Missingness Can Be an Interesting Variable in and of Itself

Missing data is often viewed as lost, an unfilled gap, but as I have demonstrated in this chapter, it is not always completely lost, given the availability **[p. 129 ↓]** of other strongly correlated variables. Going one step farther, missingness itself can be considered an outcome itself, and in some cases can be an interesting variable to explore. There is information in missingness. The act of refusing to respond or responding in and of itself might be of interest to researchers, just as quitting a job or remaining at a job can be an interesting variable. I always encourage researchers to create a *dummy variable*, representing whether a person has missing data or not on a particular variable, and do some analyses to see if anything interesting arises. Aside from attempting to determine if the data are MCAR, MAR, or MNAR, these data could yield important information.

Table 6.3 MI Estimation for MNAR-Inverse Using Weak Predictor, MCMC Estimation

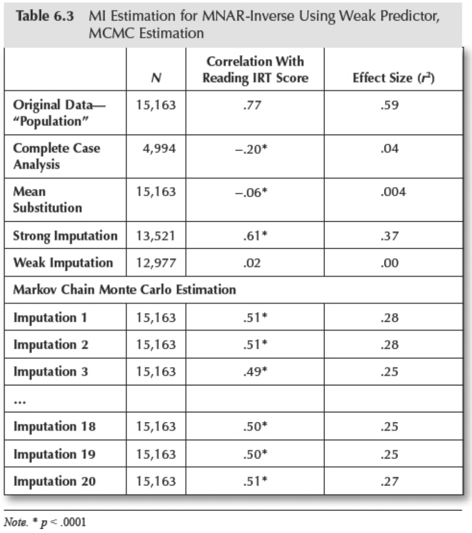


Table 6.3 MI Estimation for MNAR-Inverse Using Weak Predictor, MCMC Estimation

	N	Correlation With Reading IRT Score	Effect Size (r^2)
Original Data—“Population”	15,163	.77	.59
Complete Case Analysis	4,994	−.20*	.04
Mean Substitution	15,163	−.06*	.004
Strong Imputation	13,521	.61*	.37
Weak Imputation	12,977	.02	.00
Markov Chain Monte Carlo Estimation			
Imputation 1	15,163	.51*	.28
Imputation 2	15,163	.51*	.28
Imputation 3	15,163	.49*	.25
...			
Imputation 18	15,163	.50*	.25
Imputation 19	15,163	.50*	.25
Imputation 20	15,163	.51*	.27

Note. * $p < .0001$

[p. 130 ↓]

Imagine two educational interventions designed to improve student achievement, and further imagine that in one condition there is much higher dropout than in the other condition, and further that the students dropping out are those with the poorest performance. Not only is that important information for interpreting the results (as the differential dropout would artificially bias the results), but it might give insight into the intervention itself. Is it possible that the intervention with a strong dropout rate among those most at risk indicates that the intervention is not supporting those students well enough? Is it possible that intervention is alienating the students in some way, or it might be inappropriate for struggling students?

All of this could be important information for researchers and policymakers, but many researchers discard this potentially important information. Remember, you (or someone) worked hard to obtain your data. Do not discard anything that might be useful!

Summing Up: What are Best Practices?

This chapter ended up being a longer journey than I had intended. The more I delved into this issue, the more I found what (I thought) needed to be said, and the more examples needed to be explored. There are some very good books by some very smart people dealing solely with missing data (e.g., Little & Rubin, 1987; Schafer, 1997), and I have no wish to replicate that work here. The goal of this chapter was to convince you, the researcher, that this is a topic worthy of attention, that there are good, simple ways to deal with this issue, and that effectively dealing with the issue makes your results better.

Because we often gather data on multiple related variables, we often know (or can estimate) a good deal about the missing values. Aside from examining missingness as an outcome itself (which I strongly recommend), modern computing affords us the opportunity to fill in many of the gaps with high-quality data. This is not merely “making up data” as some early, misinformed researchers claimed. Rather, as my examples show, the act of estimating values and retaining cases in your analyses most often leads to more replicable findings as they are generally closer to the actual population values than analyses that discards those with missing data (or worse, substitutes means for the missing values). Thus, using best practices in handling missing data makes the results a better estimate of the population you are interested in. And it is surprisingly easy to do, once you know how.

[p. 131 ↓]

Thus, it is my belief that best practices in handling missing data include the following.

- First, do no harm. Use best practices and careful methodology to minimize missingness. There is no substitute for complete data[11] and some careful forethought can often save a good deal of frustration in the data analysis phase of research.
- Be transparent. Report any incidences of missing data (rates, by variable, and reasons for missingness, if possible). This can be important information to reviewers and consumers of your research and is the first step in thinking about how to effectively deal with missingness in your analyses.
- Explicitly discuss whether data are missing at random (i.e., if there are differences between individuals with incomplete and complete data). Using analyses similar to those modeled in this chapter, you can give yourself and the reader a good sense of why data might be missing and whether it is at random. That allows you, and your audience, to think carefully about whether missingness may have introduced bias into the results. I would advocate that all authors report this information in the methods section of formal research reports.
- Discuss how you as a researcher have dealt with the issue of incomplete data and the results of your intervention. A clear statement concerning this issue is simple to add to a manuscript, and it can be valuable for future consumers as they interpret your work. Be specific—if you used imputation, how was it done, and what were the results? If you deleted the data (complete case analysis) justify why.

Finally, as I mentioned in Chapter 1, I would advocate that all authors report this information in the methods section of formal research reports and that all journals and editors and conferences mandate reporting of this type. If no data is missing, state that clearly so consumers and reviewers have that important information as well.

For Further Enrichment

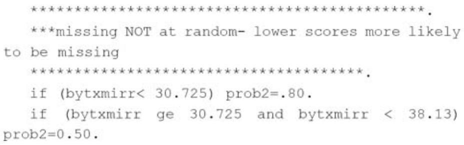
- Download from the book's website some of the missing data sets I discuss in this chapter, and see if you can replicate the results I achieved through various means. In particular, I would challenge you to attempt multiple imputation.

- Choose a data set from a previous study you conducted (or your advisor did) that had some missing data in it. Review how the missing data was handled originally. (I also have another data set online that you can play with for this purpose.)
    - Conduct a missingness analysis to see if those who failed to respond were significantly different than those who responded.
    - Use imputation or multiple imputation to deal with the missing data.
    - Replicate the original analyses to see if the conclusions changed.
    - If you found interesting results from effectively dealing with missingness, send me an e-mail letting me know. I will gather your results (anonymously) on the book's website, and may include you in future projects.
- Find a data set wherein missing data were appropriately dealt with (i.e., imputation or multiple imputation). Do the reverse of #2, above, and explore how the results change by instead deleting subjects with missing data or using mean substitution.

Appendixes

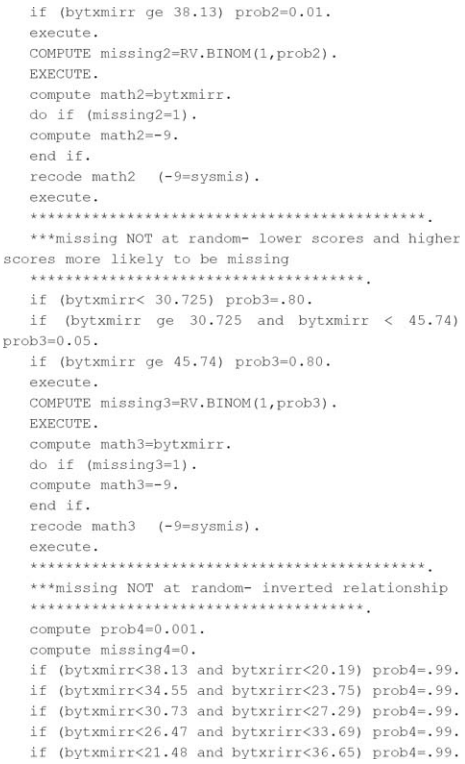
Appendix A: SPSS Syntax for Creating Example Data Sets

If you are interested in the details of how I created these various missing data sets, I am including the SPSS syntax. Also, because the MNAR-inverse data set is a particularly odd one (and one I am particularly proud of), I include scatterplots of the data points prior to and after missingness was imposed.



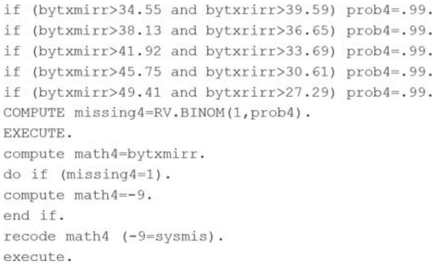
```
   ***********************************************.
   ***missing NOT at random- lower scores more likely
to be missing
   ***************************************.
   if (bytxmirr< 30.725) prob2=.80.
   if  (bytxmirr  ge  30.725  and  bytxmirr  <  38.13)
prob2=0.50.
```

[p. 133 ↓]



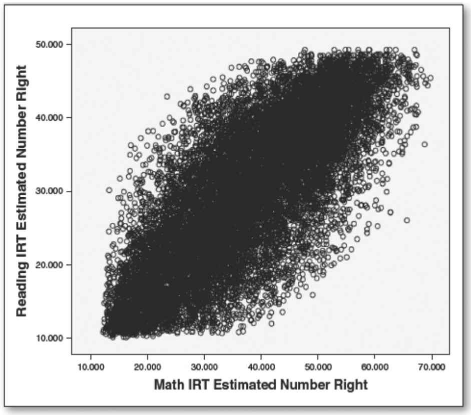
```
   if (bytxmirr ge 38.13) prob2=0.01.
   execute.
   COMPUTE missing2=RV.BINOM(1,prob2).
   EXECUTE.
   compute math2=bytxmirr.
   do if (missing2=1).
   compute math2=-9.
   end if.
   recode math2  (-9=sysmis).
   execute.
   *********************************************.
   ***missing NOT at random- lower scores and higher
scores more likely to be missing
   **************************************.
   if (bytxmirr< 30.725) prob3=.80.
   if  (bytxmirr  ge  30.725  and  bytxmirr  <  45.74)
prob3=0.05.
   if (bytxmirr ge 45.74) prob3=0.80.
   execute.
   COMPUTE missing3=RV.BINOM(1,prob3).
   EXECUTE.
   compute math3=bytxmirr.
   do if (missing3=1).
   compute math3=-9.
   end if.
   recode math3  (-9=sysmis).
   execute.
   *********************************************.
   ***missing NOT at random- inverted relationship
   **************************************.
   compute prob4=0.001.
   compute missing4=0.
   if (bytxmirr<38.13 and bytxrirr<20.19) prob4=.99.
   if (bytxmirr<34.55 and bytxrirr<23.75) prob4=.99.
   if (bytxmirr<30.73 and bytxrirr<27.29) prob4=.99.
   if (bytxmirr<26.47 and bytxrirr<33.69) prob4=.99.
   if (bytxmirr<21.48 and bytxrirr<36.65) prob4=.99.
```

[p. 134 ↓]



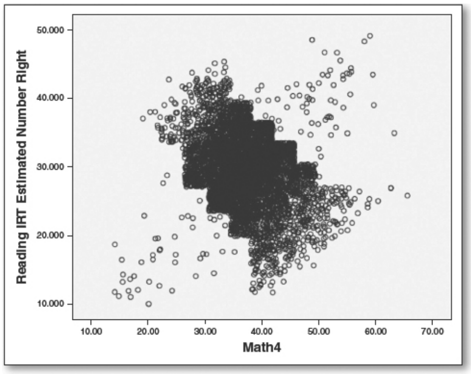
```
if (bytxmirr>34.55 and bytxrirr>39.59) prob4=.99.
if (bytxmirr>38.13 and bytxrirr>36.65) prob4=.99.
if (bytxmirr>41.92 and bytxrirr>33.69) prob4=.99.
if (bytxmirr>45.75 and bytxrirr>30.61) prob4=.99.
if (bytxmirr>49.41 and bytxrirr>27.29) prob4=.99.
COMPUTE missing4=RV.BINOM(1,prob4).
EXECUTE.
compute math4=bytxmirr.
do if (missing4=1).
compute math4=-9.
end if.
recode math4 (-9=sysmis).
execute.
```

Figure 6.3 Original Relationship Between Math and Reading Score



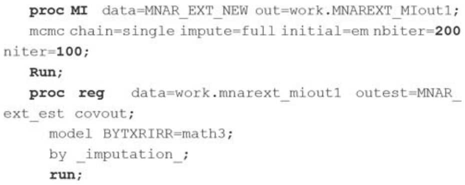
[p. 135 ↓]

Figure 6.4 Inverse Relationship Between Math and Reading Score Creating MNAR-Inverse



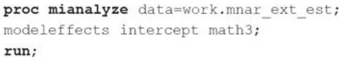
Appendix B: SAS Syntax for Performing Multiple Imputation

This SAS Syntax was used to generate multiple imputation data sets, analyze them, and report summary statistics.



```
   proc MI  data=MNAR_EXT_NEW out=work.MNAREXT_MIout1;
   mcmc chain=single impute=full initial=em nbiter=200
niter=100;
   Run;
   proc  reg   data=work.mnarext_miout1  outest=MNAR_
ext_est covout;
     model BYTXRIRR=math3;
     by _imputation_;
     run;
```

[p. 136 ↓]



```
proc mianalyze data=work.mnar_ext_est;
modeleffects intercept math3;
run;
```

Notes

1. And, of course, if this was good research, I would assume follow-up questions would ask if the respondent is in a committed, long-term relationship as well to capture the effect of being in a stable relationship with another person regardless of whether that relationship was technically an officially recognized marriage. I leave that to all the relationship researchers out there to figure out—I am just a humble quaint guy trying to help clean data.

2. Which can deal with issues like participants leaving the study (right-censored or truncated data) or entering the study at a particular point (left-censored or truncated data).

3. Once again, let us be clear that values that are “out of scope” or legitimately missing, such as nonsmokers who skip the question concerning how many cigarettes are smoked a day, are not considered missing and are not an issue (Schafer & Graham, 2002). In this example, let us imagine that non-classroom teachers (e.g., guidance counselors, teacher assistants, or other personnel) who took the initial survey were not included in the follow-up because they are not the population of interest—i.e., classroom teachers. This would be legitimate missing data.

4. Which, honestly, is darned impressive, considering how much power there was in this analysis to detect *any* effect, no matter how small.

5. This is important because, as a researcher, you would not know the true population mean, and thus would be substituting an already biased mean for the missing values.

6. Note as well that case deletion also produces artificially reduced estimates of the population standard deviation under MNAR.

7. This also is implemented relatively easily in many statistical packages. For example, the SPSS syntax command below creates an average called “average” by averaging the items if at least five of the six values are present. As mentioned in the text, this is only desirable if these items have good internal consistency.

```
      Compute average=mean.5(item01, item02,
item03, item04, item05, item06).
```

8. This is a bit of an esoteric topic to many researchers, so I will be brief and refer you to the cited references if you are interested in further information. Almost by definition, multiple regression creates an ideal fit between variables based on a particular data set. It squeezes every bit of relationship out of the data that it can. This is called *overfitting* because if you take the same equation and apply it to a different **[p. 137 ↓]** sample (e.g., if we were to predict math achievement from reading achievement and socioeconomic status in a new sample) the prediction equations are often not as accurate. Thus, the relationships in a new sample are likely to be lower, leading to “shrinkage” in the overall relationship. Thus, in the prediction literature double cross-validation is a good practice, where samples are split in two and prediction equations generated from each are validated on the other half-sample to estimate how generalizable the prediction equation is. Multiple imputation takes this to another level, essentially, by creating several different parallel analyses to see how much variability there is across samples as a function of the missing data estimation. A very sensible concept!

9. An excellent introduction and guide to this procedure and process is Yuan (2000). Though some beginning users find SAS challenging, multiple imputation through SAS is relatively painless and efficient, accomplished through only a few lines of syntax. Once programmed, the actual multiple imputation procedure that produced 20 parallel data sets, analyzed them, and reported the summary statistics took less than 60 seconds

on my laptop. For reference, I have appended the SAS syntax used to perform the first multiple imputation at the end of this chapter.

10. As the proportion of data missing increases, it is sometimes desirable to increase the number of imputed data sets to maintain a high relative efficiency. Given the ease of using SAS to create and analyze these data, and the speed of modern computers, there is little reason *not* to do so.

11. Except in certain specialized circumstances where researchers purposely administer selected questions to participants or use other advanced sampling techniques that have been advocated for in the researching of very sensitive topics.

References

Acock, A. Working with missing values. Journal of Marriage and Family, vol. 67 (2005). (4), pp. 1012–1028.

Cole, J. C. (2008). How to deal with missing data . In J. W. Osborne (Ed.), Best practices in quantitative methods (pp. 214–238). Thousand Oaks, CA: Sage.

Colebrooke, H. (1817). Algebra with arithmetic of Brahmagupta and Bhaskara . London: John Murray.

Diehl, R. (2004). The Olmecs: America's first civilization . London: Thames & Hudson.

Graham, J., Olchowski, A., and Gilreath, T. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science, vol. 8 (2007). (3), pp. 206–213.

Haitovsky, Y. Missing data in regression analysis. Journal of the Royal Statistical Society. Series B (Methodological), vol. 30 (1968). (1), pp. 67–82.

Ingels, S. Pratt, D. Rogers, J. Siegel, P. Stutts, E. Owings, J. (2004). Education Longitudinal Study of 2002: Base year data file user's manual (NCES 2004–405) .

Washington, DC: U.S. Department of Educations, National Center for Education Statistics.

Little, R. Rubin, D. (1987). Statistical analysis with missing data . New York: Wiley.

Osborne, J. W. Prediction in multiple regression. Practical Assessment, Research & Evaluation, vol. 7 (2000). (2).

Osborne, J. W. (2008). Creating valid prediction equations in multiple regression: Shrinkage, double cross-validation, and confidence intervals around prediction . In J. W. Osborne (Ed.), Best practices in quantitative methods . (pp. 299–305). Thousand Oaks, CA: Sage.

Osborne, J. W. Kocher, B. Tillman, D. (2011). Sweating the small stuff: Do authors in APA journals clean data or test assumptions (and should anyone care if they do)? Unpublished Manuscript, North Carolina State University.

Raudenbush, S. W. Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (Vol. 1). Thousand Oaks, CA: Sage.

Rubin, D. Inference and missing data. Biometrika, vol. 63 (1976). (3), pp. 581–592.

Schafer, J. (1997). Analysis of incomplete multivariate data . London: Chapman & Hall/CRC.

Schafer, J. Multiple imputation: A primer. Statistical Methods in Medical Research, vol. 8 (1999). (1), pp. 3–15.

Schafer, J. and Graham, J. Missing data: Our view of the state of the art. Psychological Methods, vol. 7 (2002). (2), pp. 147–177.

Stuart, E. A., Azur, M., Frangakis, C., and Leaf, P. Multiple imputation with large data sets: A case study of the children's mental health initiative. American Journal of Epidemiology, vol. 169 (2009). (9), pp. 1133–1139.

Yuan, Y. (2000). Multiple imputation for missing data: Concepts and new development . Paper presented at the Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, Cary, N.C. Retrieved from http://support.sas.com/rnd/app/papers/multipleimputation.pdf

http://dx.doi.org/10.4135/9781452269948.n6