26

Missing Data

If we have data, let's look at data. If all we have are opinions, let's go with mine

- JIM BARKSDALE

INTRODUCTION

This chapter considers strategies for dealing with missing data in RETs. I begin by stating the obvious: The best strategy for dealing with missing data is not to have any. Although I say this somewhat tongue-in-cheek, it often is possible to adopt practices that discourage missing data. One reason people do not answer sensitive questions is because of embarrassment or because they think questions are too personal. You can alleviate such concerns by avoiding face-to-face respondent-interviewer interactions (e.g., by having answers completed on a self-administered form or a computer) and by using instructional sets to motivate the honest answering of questions (see Chapter 3). Keeping surveys short and concise also can help. If people drop out of a longitudinal study, sometimes they are amenable to completing a few measures central to the study at later waves if they are fairly compensated and response burden is minimized. If it is possible that you will lose contact with a participant in a longitudinal study, there are well-established methods you can use to maintain contact (Madden et al., 2017).

In RETs, people sometimes drop out of studies during treatment without fully completing the treatment. I refer to these individuals as **treatment dropouts**. They often are a source of missing data. Other individuals complete treatment but do not complete one or more of the posttreatment surveys for purposes of assessing treatment response. I refer to these individuals as being **lost to follow-up**. Although treatment dropouts typically also are lost to follow-up, they do not have to be; they still can complete posttreatment surveys. The issues that need to be addressed for missing data in RETs are somewhat different for those who complete treatment but who are lost to follow-up versus those who are treatment dropouts. My focus in this chapter is on how we handle missing data analysis for those lost to follow-up. These individuals have completed treatment per protocol but they have, for one reason or another, missed an assessment at the posttest or a follow-up. In Chapter 27, the analytic problem shifts to treatment dropouts, requiring a different approach than those lost to follow-up. Not only must I take into account that dropouts may miss an assessment session and have missing data because of this, but I also must account for the fact that dropouts have been exposed to

some fraction of the intervention.

My focus in this chapter is on missing data due to loss to follow-up in regression and SEM analytic contexts. I first describe traditional methods that have historically dominated how missing data are handled. I then describe three mechanisms that cause missing data and that impact how you deal with missing data. Next, I discuss modern methods for working with missing data, including full information maximum likelihood (FIML) methods, full information Bayesian approaches, single imputation methods, and multiple imputation methods. The number of strategies for addressing missing data is considerable and choosing between them can be both complex and confusing. I make recommendations for you to consider. Like many of the chapters in this book, one could write an entire text on the topic of missing data so I am selective in what I cover. My goal is to provide you with overviews of key issues to help you contextualize missing data in RETs. Despite this more cursory orientation, the chapter is long and it may take you several "sittings" to get through it.

TRADTIONAL APPROACHES TO MISSING DATA

Two of the more popular strategies for dealing with missing data are **listwise deletion** and **pairwise deletion**. Suppose you plan to conduct a multiple regression analysis predicting Y from X and Z but some of your cases have missing data on one or more of the variables. With listwise deletion, if a case is missing data on any one of the variables, the case is eliminated completely from the analysis. This strategy has several disadvantages. First, you can end up deleting a large number of cases which can drastically reduce your effective sample size. The result is lower statistical power and wider confidence intervals. For example, consider a data set of 10 variables with N=200. There are a total of 2,000 observations (10 X 200 = 2,000) in the data matrix. Suppose that 10% of the 2,000 observations are missing. If all of the missingness occurs on the same, say, 25 individuals, the sample size loss in terms of N will only be 25, so the listwise deleted sample will have N = 175. However, it is possible that the missingness is spread out across more individuals than this to the point that the listwise deleted N becomes unacceptably small. Another limitation of listwise deletion is that you ignore the information that was provided by the listwise deleted person on the variables s/he responded to. If a person is missing information on Y, but provides information on X and Z, why not take the information you have on X and Z into account when calculating, say, a mean for X, a mean for Z, and a covariance between X and Z? Unless including such information introduces non-trivial bias, it makes sense to use it.

With pairwise deletion, you also delete a case with missing data, but only for

calculations involving the specific variable where the data are missing. If a person has scores on Y and Z but not X, you would include the person when calculating, for example, the correlation between Y and Z, but omit the person when calculating the correlation between X and Z. Multiple regression requires you estimate the covariance matrix between all variables in the model, so some of the covariances will be based on different sample sizes with pairwise deletion. A dilemma then becomes what sample size to use when defining standard errors for the regression/path coefficients. Many computer programs that allow for pairwise deletion use the smallest N on which a statistic was based, but this is an ad hoc approach. Another problem with pairwise deletion is that it is possible to obtain patterns of correlations that are theoretically impossible. For example, given the values of the correlation between X and Y and X and Z, one can calculate the range of values that the correlation between Y and Z must fall within. With pairwise deletion, it is sometimes possible to obtain correlations outside this range.

If there is no meaningful bias in the sources of missing data and if missing data are minimal, then pairwise or listwise deletion of data usually will not be problematic. But as the amount of missing data becomes more substantial, these methods can be unsatisfactory. Another strategy that is sometimes used is **mean imputation**. This calculates the mean of a variable for cases where data are present and then substitutes the mean score for the cases where data are missing on that variable. This is a questionable strategy. It often produces biased estimates of standard deviations, covariances, and correlations. For example, for annual income if I have 20% missing cases and I impute the mean value of the observed scores, say $20,000, to the missing values, then I will be imputing for 20% of the people the same value of $20,000, which lowers the standard deviation of the scores. The artificially lowered standard deviation will, in turn, affect estimated standard errors, p values, and confidence intervals.

MISSING DATA MECHANISMS

Statisticians distinguish three types of missingness. Data can be **missing completely at random** (MCAR), **missing at random** (MAR), or **missing not at random** (MNAR, also called **non-ignorable missingness**). The distinctions between the mechanisms are subtle and sometimes mischaracterized. I use a simple example outside the context of an RET to explain the differences. The concepts readily extend to RETs.

As is well known, there are sex differences in the weight of young adults with males tending to weigh more than females. Suppose I want to estimate the coefficient for the effect of biological sex on weight to determine the magnitude of the weight difference between males and females. The underlying causal model using sample notation is

$$\text{Weight}^* = a + p_1 \text{ sex} + d_1 \quad [26.1]$$

where biological sex is dummy coded 0 = female, 1 = male, a is the intercept, and p_1 is the unstandardized path/regression coefficient. Suppose in my study, interviewers ask young adults for consent to measure their weight on a highly accurate scale and, given consent, obtain a score for weight in pounds for the individual. Suppose I have complete data for an index of whether a study participant is male or female but missing data on the weight variable due to some people not consenting to be weighed. I can distinguish two versions of the weight variable. Weight* refers to the weight of each person including the weight of people I was unable to collect weight data on. This variable is hypothetical because, in reality, I do not have access to weight information for everyone; but I assume this variable, in principle, exists. By contrast, the variable Weight refers to the weight data I collect for all individuals but with a missing value (e.g., an NA, a dash, or a blank) substituted for individuals whose weight I am unable to measure. This variable has only partial weight information. I can specify a third variable for each individual called Response (R) that is scored 0 = I did not obtain weight data for the person or 1 = I obtained weight data for the person. This is a dummy indicator for missingness.

In theory, I can calculate the probability of missingness occurring for different profiles of the predictor variable, i.e., I can specify the probability of missing data occurring for males and I can also calculate it for females. If the probability of missingness is the same for males and females, then there is no association between biological sex and R, or stated more technically, the probability of R is conditionally independent of biological sex. If, however, females are less likely to agree to have their weight measured than males, then the probability of R is not conditionally independent of biological sex. I also can ask, in theory, if the probability of missingness is associated with Weight* or, stated more formally, whether the probability of R is conditionally independent of the value of Weight*. Suppose that people who are overweight are less likely to agree to having their weight measured. This means that the probability of R is not independent of Weight*; that the two constructs are associated.

The presence of these two types of independence characterize one type of missing data, namely data that are **missing completely at random**. For data to be MCAR, (a) the probability of missingness (R) should be independent of the predictor(s) in our model, in this case, biological sex and (b) the probability of missingness should be independent of the true values of the outcome, in this case Weight*. Indeed, MCAR requires that missingness be unrelated to all variables because it is what we typically think of when we think of haphazard missingness. The MCAR assumption has dominated social science research approaches to missing data until relatively recently.

Either of the two independence properties of MCAR noted above can be violated

due to a variety of phenomena. For example, Weight* might be associated with R because Weight* and R share a common cause, i.e., there is a confounding influence on both variables. Social class might influence Weight* and it might also influence the probability of failing to consent to having one's weight measured, producing an association between R and Weight*. Alternatively, Weight* might be associated with R not because it shares a common cause with R but instead because Weight* directly influences R because overweight people are less likely to give consent to be weighed.

Figure 26.1a shows an influence diagram of the effect of biological sex on Weight* whose path coefficient (p_1) I seek to estimate. Figure 26.1b adds to this influence diagram the observed variable Weight, which contains missing data. The model states that Weight is a causal function of R and Weight*. Note that the function is complex and non-linear, so I omit traditional path coefficient symbols as a means of acknowledging this. The function that expresses Weight as a function of R and Weight* is:

$$\text{Weight} = f(\text{R, Weight*}) = \begin{Bmatrix} \text{Weight*} & \text{if R} = 1 \\ \text{NA} & \text{if R} = 0 \end{Bmatrix}$$

which is just a succinct way of saying the following: Weight equals the value of Weight* for a given individual if R = 1 for that individual, i.e., if I am able to weigh the person; if R = 0 for the individual, Weight is set to NA. In Figure 26.1b, both sex and Weight* are unrelated to R because there are no arrows connecting R to them. The data are MCAR. Rubin (1976) has shown that if I regress Weight onto biological sex in such a scenario using listwise deletion or one of the other missing data strategies I discuss below, I should obtain an unbiased estimate of the population value of p_1 in Figure 26.1a.

Figure 26.1c alters Figure 26.1b to introduce a causal link between sex and the probability of R, i.e., females, perhaps because of a greater concern for body image, are less likely to consent to being weighed than males. The dependency introduced by p_2 violates MCAR requirements. However, the situation is salvageable analytically if I can assume that Weight* is not associated with the probability of R when just males are considered and that Weight* also is not associated with the probability of R when just females are considered. If this is true, then the relationship between Weight* and the probability of R vanishes when I condition the probability of R on sex, i.e., when I hold sex constant. This yields what is known as data that are missing at random (MAR), i.e., the required independence properties for valid estimation hold *as long as I explicitly hold constant the predictor that is a determinant of missingness on the target variable*. Rubin (1976) also shows that in this scenario, if I regress Weight onto my predictor, biological sex, then most of the missing data strategies I discuss below (but not listwise deletion) should again yield an unbiased estimate of the population value of p_1 in Figure 26.1a.

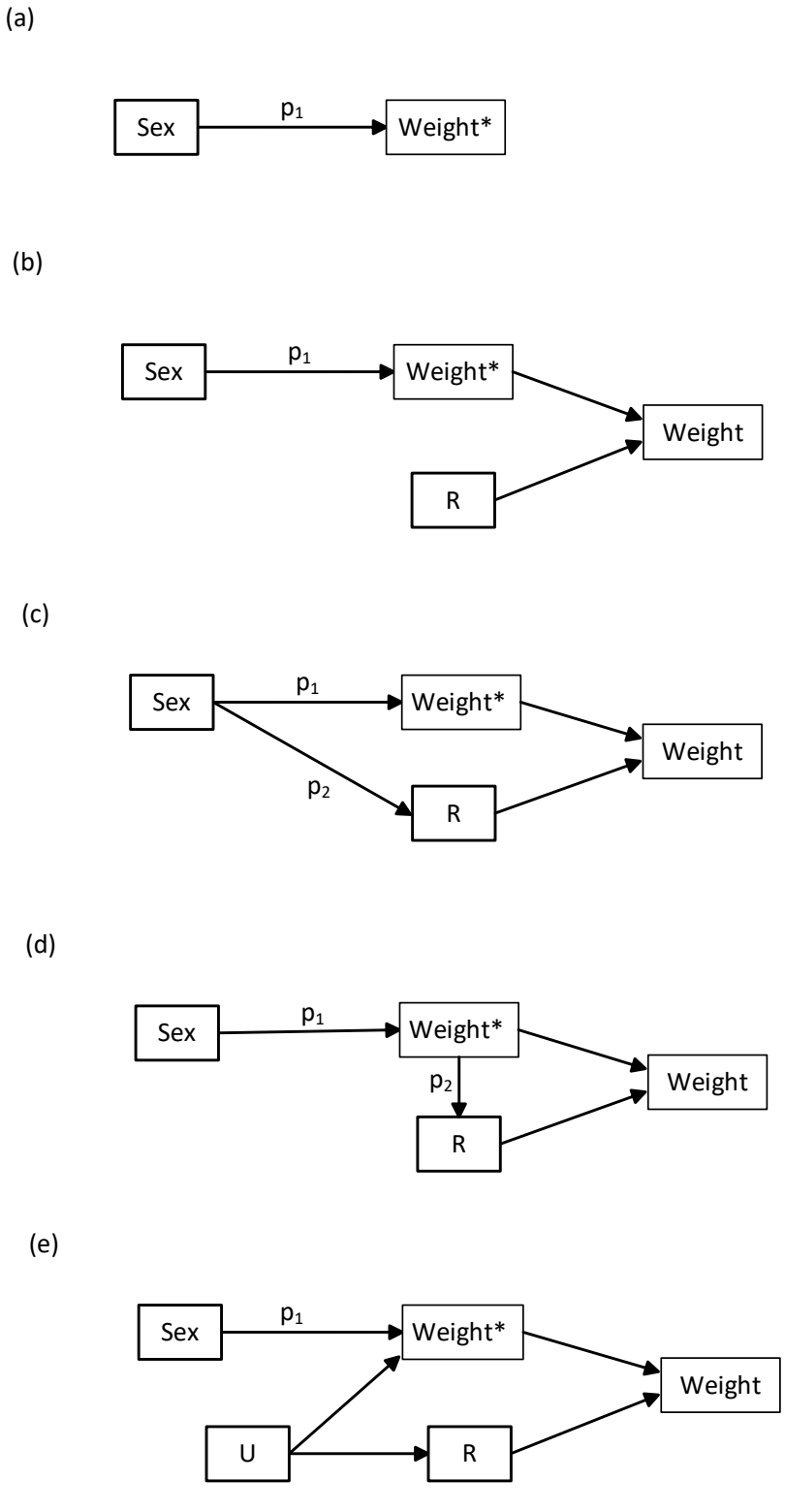


FIGURE 26.1. Mechanisms of missingness

Figure 26.1d presents a more insidious case where Weight* directly influences the probability of R. In this case, the properties of both MCAR and MAR are violated because even when I hold biological sex constant, Weight* remains related to R. Conditioning on biological sex does not help matters. The data are MNAR. It turns out I still can obtain unbiased estimates of p_1 in Figure 26.1a but doing so is challenging. I discuss approaches to dealing with missing data that are MNAR later in this chapter.

Finally, Figure 26.1e shows a case where the dependency between Weight* and the probability of R is due to an unmeasured confounding variable, U. Because U is unmeasured, I can't control for it statistically to remove the dependency it causes between R and Weight*, an analytic strategy I relied on for the case of MAR with respect to biological sex. Because of this, the data are MNAR and, again, I must address it using specialized methods. Whenever there is an unmeasured variable causing violation of the missing data conditional independence assumption, the data are MNAR.

When we analyze studies with missing data, we need to invoke a theory of missingness to determine how best to handle it. We need to think about whether the data are likely MCAR. If we think the data are not but that they are MAR, then we need to use a missing data analytic strategy that can accommodate MAR to remove the offending dependencies that result in biased estimates. If we think the data are MNAR, then we need to adopt analytic strategies to estimate parameters under MNAR. There are many ways that the dependencies described above can arise, so developing a theory of missingness is important. The situation becomes even more complicated when there are multiple predictors with complex relationships among them and those predictors also have missing data (see Allison 2017). In the final analysis, large amounts of missing data can make life difficult for researchers. It is better to keep it to a minimum.

ASSESSMENT OF BIAS IN MISSING DATA

One cannot easily empirically determine if data are MCAR or MAR for a variable, Y, because to do so one needs information on the respondents whose scores on Y are missing. How can we know if people who score low (or high) on Y are more (or less) likely to have missing Y data if we do not know their Y scores? The answer is that we can't. There is no unambiguous test to assert that missing data are MCAR or MAR.

To gain perspectives on whether data are MCAR one can examine associations between a missing data dummy variable for Y and other variables in one's model or the data set more broadly. A **missing data dummy variable** for Y assigns a person a score of 1 if the person has missing data on Y, otherwise a score of 0. There should be no meaningful associations between this dummy variable and other variables in the data set because MCAR assumes data are missing completely at random, i.e., that it is haphazard.

If there are no such associations, then this increases confidence that the data are indeed MCAR. However, unless one can build a compelling case that missing data on Y also are unrelated to scores on Y*, then assumptions of MCAR (and MAR) must be tentative. For a more in depth discussion of this matter, see Raykov (2011) and Raykov et al. (2012).

A seldom used strategy to evaluate if missing data on Y are related to scores on Y is to correlate a missing data dummy variable for Y with a proxy measure of Y if a proxy is available. For example, a researcher might administer two interchangeable measures of depression, Measure A and Measure B. The researcher can create a missing data dummy variable for Measure A and then correlate it with values on Measure B. The correlation should be trivial. This strategy, of course, requires that the proxy measure lacks missing data or that its missing data is randomly dispersed, neither of which may be the case.

Many software packages offer a multivariate version of the "dummy variable" test for MCAR. This test essentially creates missing data dummy variables for each variable in one's model and then simultaneously tests the associations between the dummy variables and other variables considered as a collective. The test is called **Little's MCAR test** and it should be statistically non-significant if there is no bias. Keep in mind, however, that the name of the test is a misnomer because, technically, MCAR also requires that missingness on Y not be associated with scores on Y*, which is not addressed in the test. Also, low statistical power will tend to produce non-significant results even if bias is present. A significant result on Little's tests leads one to doubt MCAR; a non-significant result does not allow one to make a definitive conclusion, but it can increase one's confidence that the data are MCAR if the test is adequately powered.

MISSING AT RANDOM IS A MATTER OF DEGREE

The literature on missing data is replete with "all-or-none" statements about missing data strategy appropriateness. For example, it often is stated that listwise deletion strategies are only appropriate when data are MCAR but not when data are MAR. Such statements ignore, among other things, the fact that the properties of MCAR and MAR are a matter of degree. Data may be MAR rather than MCAR but may deviate from MCAR to such a small extent that the deviation has little impact on inferences for methods that assume MCAR. In the example for the effect of sex differences on weight, it may be the case that individuals with higher weight are indeed more likely to have missing data. However, if the path coefficient in a linear probability model linking the true weight in pounds to the probability of missing data is only 0.0001, then the probability of missing data increases by only 0.01 for every 100 pound increase in weight. It is unlikely listwise deletion strategies will produce biased parameter estimates in this case; the *degree* of departure from MCAR must be consequential.

MISSING DATA BIAS IS NOT ALWAYS BAD

Sometimes we find that missingness is associated with some other variable in our data set but the bias is not relevant to the research question being addressed. Not all bias matters. Suppose I want to estimate the percent of people who have been divorced in a population and, unbeknownst to me, the divorce rate is 33%. Suppose the divorce rate is the same for lower, middle, and upper class individuals, i.e., it is 33% across the SES spectrum. Suppose further that lower class individuals are less likely to answer a question about divorce status on a survey, i.e., there is missing data bias on the divorce variable as a function of income. I collect data from a random sample of 500 individuals and use listwise deletion by eliminating individuals that did not answer my question about divorce. It turns out this listwise deleted sample will be biased in terms of SES because I have unwittingly eliminated more lower class than upper class individuals given lower class people are less likely to answer the question about being divorced. Despite this, the bias in the listwise deleted sample will not affect my estimate of the population divorce rate; it will still be 33% (plus sampling error) because the divorce rate is the same across the class spectrum. The listwise sample *is* representative of the population *on variables that matter* for the question I seek to answer. My general point is that although samples can be biased, the bias may not matter.

The issue of missing data bias becomes further muddled when one takes into account the way social science research is conducted in practice. As discussed in Chapter 4, the traditional way of thinking about populations and samples involves two steps. First, we define the population of individuals we want to make statements about. Second, we enact procedures that generate a random (or approximately random) sample from that population. However, it is possible to turn this logic on its head by reversing the process; A researcher conducts a study on a group of individuals and then declares that the group of individuals represents a random sample from some unspecified population. The researcher then seeks to specify the population the sample likely is randomly selected from. We are still dealing with a population and a random sample from that population. It is just that we are using the sample to drive the specification of the population that the sample can be construed as a random sample from. The former approach is called a **population-then-sample** approach and the second is called a **sample-then-population** approach. For elaboration, see Chapter 4.

In practice, the sample-then-population strategy is far more common in social science research than the population-then-sample approach, especially for RCTs and RETs. In this sense, sample bias due to missing data can be construed as just one of many other factors that researchers must take into account when trying to determine or argue for the population their sample represents. When using listwise deletion, for example, the

population you construe your sample as a random sample from is, among other things, people who choose to respond to questions on your survey or who provide complete data for the variables in the model you are testing. How limiting is this condition to your conclusions? If you are studying a fundamental mental, social, or biological process that likely generalizes across a broad spectrum of individuals, does it really affect your conclusions if in your sample, low SES individuals are more likely to drop out of the study? To take an extreme example, if I want to determine the color of blood, does it really matter if my sample to make that determination is under-represented by people who are low in SES? Just as we often include disclaimers in Discussion sections about generalizing to populations with demographic characteristics different than people in our study, one also can make such disclaimers to those who provide complete data if one uses the listwise deletion of missing data strategy. In the sex differences in weight example described earlier, if you use listwise deletion, you might limit conclusions to a population of individuals who assent to having their weight measured. One then judges the meaningfulness and limitations of the research in light of this.

In sum, matters of missing data often are characterized in the literature as "black and white" when, in fact, there are gray areas. I again encourage you as a first priority to adopt methodological practices that avoid missingness. Faced with less than satisfactory amounts of missing data, you need to think long and hard about your theory of missing data, i.e., the sources and correlates of missing data, and then let that theory guide the analytic decisions you make. You may identify systematic missing data tendencies that do not matter for the questions you seek to answer, which are then non-problematic. If you use a sample-then-population framework, then your generalizations are constrained by the particular missing data strategy you adopt but this may or may not be problematic for your research goals. Missing data dynamics are nuanced, not all or none despite statements by some to the contrary (e.g., Newman, 2014).

PATTERNS OF MISSING DATA

When analyzing data, it is useful to have a sense of the patterns of missing data for the variables in your model. Mplus offers two approaches for doing so. One approach is to perform a separate multivariate analysis of the variables in your model that uses the `BASIC` analysis structure in Mplus. Example syntax for an Ret example is in Table 26.1.

Table 26.1: Mplus Syntax for Missing Data Patterns

```
1. TITLE: MISSING DATA PATTERNS ;
2. DATA:   FILE = missingexample1.dat ;
```

```
3. VARIABLE:
4. NAMES = id treat days educ income m1 m2 m3 y y2
6. x1 x2 x3 x4 x5 x6 x7 x8 x9 ;
7. USEVARIABLES = treat educ income m1 m2 y ;
8. MISSING = ALL (-9999) ;
9. ANALYSIS:  TYPE = BASIC ;
```

Lines 1 through 8 are standard Mplus syntax. Line 9 tells Mplus to use the `BASIC` option. There is no model to specify as this option by default calculates only the means, variances, and covariances of the variables as well as missing data information. You do not need to specify an `OUTPUT` line because the program has a default `OUTPUT` structure. In this program, I seek to isolate multivariate missing data patterns for an outcome variable (Y), a treatment variable, two covariates (maternal education and income) and two mediators (M1 and M2). They are specified in Line 7.

The patterns output is reported in the form of a matrix and lists different pattern numbers as columns and each variable name in your model as a row. An `x` in a cell indicates that respondents provided data on the row variable, like this

```
    MISSING DATA PATTERNS (x = not missing)

           1  2  3  4  5  6  7  8  9 10 11 12
TREAT      x  x  x  x  x  x  x  x  x  x  x  x
EDUC       x  x  x  x  x  x  x  x  x
INCOME     x  x  x  x  x  x              x
M1         x  x  x  x           x  x     x  x  x
M2         x  x           x  x  x  x  x  x  x  x
Y          x     x        x     x     x  x  x
```

```
    MISSING DATA PATTERN FREQUENCIES

   Pattern   Frequency     Pattern   Frequency     Pattern   Frequency
         1         171           5          18           9           3
         2          45           6           9          10           3
         3          15           7           6          11           3
         4          15           8           6          12          18
```

Pattern 1 (the first column) represents individuals who provided complete data. There were 171 such individuals. Pattern 2 represents individuals who provided complete data on all variables except Y. There were 45 such individuals. And so on. Some strategies for dealing with missing data take the different patterns into account, as I describe later.

A second way of examining missing data patterns in Mplus is to add the keyword `PATTERNS` to the `OUTPUT` line of the Mplus program you are executing to formally test your model. This option can produce different results than the above approach. The Mplus default when modeling continuous outcomes is to listwise delete cases with

missing data on exogenous variables but to use a modern missing data technique known as full information maximum likelihood (FIML) for endogenous variables.[1] The patterns shown by specifying the PATTERN keyword during formal modeling are defined *after* the cases with missing data on the exogenous variables have been listwise deleted. By definition, there will be no missing data on your exogenous variables because cases on those variables with missing data have been deleted for purposes of model testing.[2]

In addition to missing data patterns, Mplus outputs relevant variance/covariance coverage information. Here is what the output looks like:

```
   PROPORTION OF DATA PRESENT

              TREAT         EDUC          INCOME        M1            M2
              ________      ________      ________      ________      ________
 TREAT           1.000
 EDUC            0.923         0.923
 INCOME          0.885         0.875         0.885
 M1              0.904         0.827         0.798         0.904
 M2              0.904         0.827         0.788         0.808         0.904
 Y               0.702         0.683         0.663         0.635         0.654

            Covariance Coverage
               Y
               ________
 Y               0.702
```

Each entry specifies the proportion of cases that did *not* have missing data for each variance and covariance in the input covariance matrix. The diagonals of the matrix are the variances. If a variable has 1.00 in its diagonal, this means there was no missing data on that variable. This was the case for the variable TREAT. Variable EDUC had 92.3% of cases with data and 100.0 - 92.3 = 7.7% cases with missing data. Variable M1 had 90.4% of cases with data and 100.0 – 90.4 = 9.6% cases with missing data. The off-diagonals provide the proportion of cases that had complete information on the two referenced variables. For the variables EDUC and INCOME, 87.5% of the sample provided information on both variables. For the variables M1 and Y, 63.5% of the sample provided information on both variables; 100.0 - 63.5 = 36.5% had missing data on at least one of them.

Mplus also can be used to construct missing data dummy variables and correlate them with other variables to explore if there are biases associated with missing data. Here is the Mplus code that accomplishes the task for the current example:

[1] I describe FIML later in this chapter
[2] There are exceptions, which I elaborate on later.

```
1. TITLE: MISSING DATA DUMMY VARIABLES ;
2. DATA: FILE = missingexample1.dat ;
3. VARIABLE:
4. NAMES ARE id treat educ income m1 m2 y  ;
5. USEVARIABLES ARE treat educ income m1 m2 y
6. deduc dincome dm1 dm2 dy  ;
7. MISSING = ALL (-9999) ;
8. DATA MISSING:
9. NAMES = educ income m1 m2 y ;
10. TYPE = MISSING ;
11. BINARY = deduc dincome dm1 dm2 dy ;
12. ANALYSIS: TYPE = BASIC ;
```

Lines 1 to 4 follow standard Mplus programming. Line 5 specifies the variables to focus the bias analysis on but adds at the end a list of variable names of my choosing (that I placed on Line 6) for the missing value dummy variables to be created; each will be scored for each individua in the data set 1 = data is missing for the variable, 0 = data is not missing for the variable. I omit a dummy variable for the `TREAT` variable because there are no missing data on it; its dummy variable would be a constant of 1s across all respondents and would correlate zero with all variables. Line 7 tells Mplus the missing data value for the input variables, in this case -9999. The `DATA MISSING` command on Line 8 tells Mplus I want to conduct specialized missing data analyses. Line 9 provides the names of the variables I want to create dummy variables for. Note that I again omit the `TREAT` variable given no missing data on it. The `TYPE` command on Line 10 tells Mplus I want to use standard missing data analysis (M plus offers two other options, which I do not consider here) and Line 11 tells Mplus to create binary missing data dummy variables whose names are listed to the right. The order of these names map onto the order in Line 9. Here is the output for correlations between the variables:

```
           Correlations
              TREAT         EDUC          INCOME        M1            M2
              _____         _____         _____         _____         _____
 TREAT          1.000
 EDUC          -0.001         1.000
 INCOME         0.122         0.459         1.000
 M1             0.141         0.152         0.163         1.000
 M2            -0.033        -0.113         0.067         0.341         1.000
 Y              0.108         0.134         0.085         0.171         0.080
 DEDUC          0.006        -0.105        -0.297         0.138        -0.073
 DINCOME        0.007        -0.165        -0.044         0.042        -0.025
 DM1            0.006        -0.124        -0.041         0.002         0.027
 DM2            0.006         0.082         0.062         0.056         0.004
 DY             0.034        -0.058        -0.083        -0.092        -0.035
```

Correlations	Y	DEDUC	DINCOME	DM1	DM2
Y	1.000				
DEDUC	-0.089	1.000			
DINCOME	-0.294	0.686	1.000		
DM1	-0.078	-0.094	-0.016	1.000	
DM2	0.226	-0.094	-0.118	-0.106	1.000
DY	-0.004	0.285	0.291	0.001	0.144

As examples, the correlation between the treatment condition `TREAT` and missingness on maternal education `DEDUC` was 0.006. The correlation between `INCOME` and missingness on `DM1` was -0.044. The correlation between missingness on education `DEDUC` and missingness on income `DINCOME` was 0.686. And so on. In general, it is ideal if we observe low correlations between the missingness dummy variables and the target variables in the model. This generally was the case, with a few exceptions; for example, the correlation between Y and missingness on income `DINCOME` was -0.294 such that individuals with higher values of Y were less likely to report their income. Several of the strategies for dealing with missing data that I discuss below can accommodate bias introduced by this dependency, i.e., they accommodate data that are MAR.

A NUMERICAL EXAMPLE

I illustrate the major approaches to missing data using a numerical example that has complete data. I later introduce missing data into the data consistent with an MCAR mechanism and then compare the results I get to the results for the complete data analysis for different missing data analytic methods. The study is an RET that evaluates an intervention to reduce pain for people living with chronic pain (N = 300). The outcome variable is the Brief Pain Inventory (BPI) which ranges from 0 to 10 with higher scores indicating more disabling chronic pain. Rough verbal descriptors associated with the scale scores are no pain = 0, mild pain = 1-3, moderate pain = 4-6, severe pain = 7-10. The scale was administered at baseline and at 6 months after the intervention.

The intervention addressed three mediators, the learning of relaxation skills to minimize pain, increasing knowledge about effective strategies for taking opioid pain medications, and restructuring pain cognitions, i.e., teaching people how to think about pain differently. Measures of each mediator were obtained at baseline and post intervention, which lasted a total of 10 weeks. Mediator scores could range from 0 to 10, with higher scores indicating more positive standing on the mediator. Biological sex (0= female, 1 = male) was included as a baseline covariate for the outcome. All quantitative variables were generated to be approximately normally distributed in the population.

The model is in Figure 26.2 (I omit covariates from the figure to reduce clutter). For data that were complete, the model provided reasonable data fit. The chi square test of perfect model fit in the population was statistically non-significant (chi square = 17.70, df = 18, p < 0.48); the RMSEA was <0.001 with an 90% upper confidence limit of .05 and a p value of close fit of 0.99; the CFI was 1.00 and the standardized RMR was 0.030. There were no statistically significant tests of the difference between elements of the predicted and observed covariance matrices and no meaningful modification indices greater than 4.

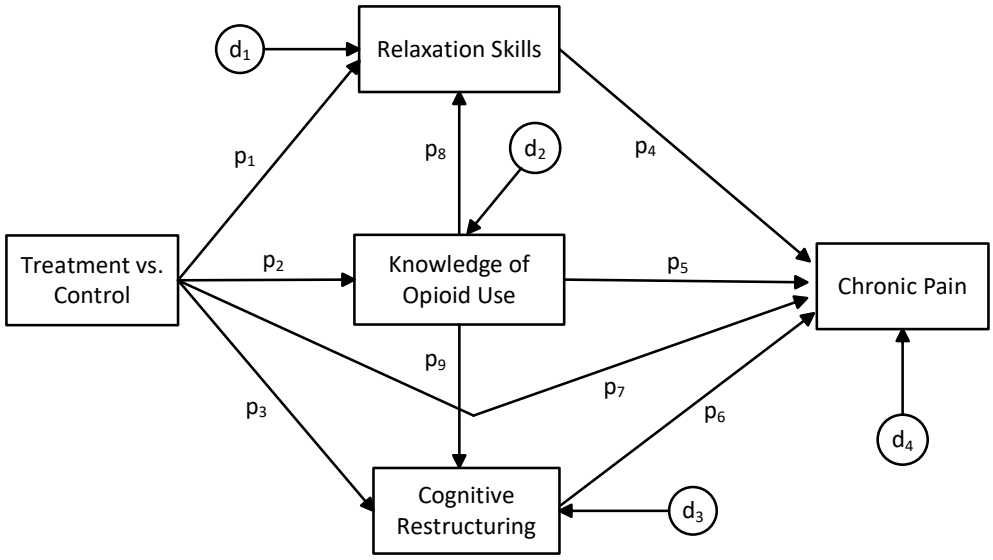


FIGURE 26.2. Chronic pain example

In addition to a sample with no missing data (which I will refer to as the complete case data set), I created a data set with missing data by using the same data as the complete case but where I changed some of the observed values to missing values. In this new data set, I had no missing data for the baseline (exogenous) variables except for biological sex, where I introduced missing values for four randomly selected individuals. For each of the three mediators measured at the immediate posttest, I introduced 3% missing data with the cases that I changed to missing values being randomly determined. Thus, there was 3% missing data for relaxation skills (RS), 3% missing data for knowledge about opioid medications (KO) and 3% missing data for cognitive restructuring (CR) when measured at the immediate posttest. The cases I changed to missing values on one mediator did not necessarily have missing data on another mediator – the random selection for creating missing data was independent across the

three mediators. For chronic pain, which was measured 6 months after treatment completion, I changed 15% of the cases to missing values, again randomly. The data are, in essence, missing completely at random (MCAR). If I listwise delete cases with missing data, the final N drops from 300 to 231, a decline of 23%. These rates are not that different from what one might observe in a longitudinal RET in which there are no treatment dropouts (again, I treat the case of treatment dropouts in Chapter 27), the missing data at the immediate posttest for the mediators is due to small amounts of non-response to items on the assessment battery, and there is a higher missing rate for chronic pain (CP) because of loss to follow-up as people return to their normal lives for the ensuing six months.

Table 26.2 presents the Mplus syntax to test the model in Figure 26.2 but it also includes the baseline covariates as is typical for RETs. The programming follows conventions from prior chapters, so there is no need to comment on it.

Table 26.2: Mplus Syntax for Numerical Example

```
1. TITLE: ANALYSIS OF CHRONIC PAIN ;
2. DATA:   FILE = missingexample2a.dat ;
3. VARIABLE:
4. NAMES = ID RS KO CR CP BRS BKO BCR BCP T BSEX ;
5. ! variables beginning with B are baseline covariates
6. ! T is treatment condition
7. USEVARIABLES RS KO CR CP BRS BKO BCR BCP T BSEX ;
8. MISSING ARE ALL (-9999) ;
9. ANALYSIS:
10. ESTIMATOR = MLR ;
11. MODEL:
12. RS ON BRS T ;
13. KO ON BKO T ;
14. CR ON BCR T ;
15. CP ON BSEX BCP RS KO CR T ;
16. MODEL INDIRECT:
17. CP IND T ;
18. OUTPUT:
19. SAMP STANDARDIZED(STDYX) MOD(ALL 4) RESIDUAL CINTERVAL PATTERNS TECH4 ;
```

MODERN STRATEGIES FOR DEALING WITH MISSING DATA

Four general approaches to dealing with missing data are described here: Maximum likelihood, Bayesian full information estimation, principled single imputation, and multiple imputation. I discuss each, in turn.

Maximum Likelihood Approaches

Maximum likelihood strategies for missing data assume data are MCAR or MAR. There are different maximum likelihood strategies for addressing missing data. One is called the expectation-maximization method, also known as the **EM method**. I do not delve into the particulars of it here; it is described by Little and Rubin (1987) and Allison (2001). A different but common maximum likelihood strategy is called **full information maximum likelihood** (FIML) missing data analysis or **direct maximum likelihood** missing data analysis. FIML typically provides less biased standard errors than the EM method (but see Enders & Peugh, 2004, for qualifications) so it tends to be preferred to EM. FIML is available in most structural equation modeling (SEM) software. It is the default method for treating missing data for endogenous variables in Mplus; see my website for links on how to implement FIML in lavaan. The fundamentals of FIML are described in Enders (2022). FIML does not involve imputation of scores for missing data, although some people mistakenly think it does so. Rather, it seeks to estimate values of the population covariances, variances and means based on the available information.

The formal mathematics of FIML in SEM contexts are too complex to develop here but I will try to give you an intuitive sense of how it operates. However, I first need to digress somewhat into the basics of maximum likelihood estimation more generally.

Maximum Likelihood and Log Likelihoods

One principle we learn in introductory statistics is that if a set of scores is normally distributed, then the distribution of those scores has certain properties. We know the distribution is symmetrical, that it is bell-shaped, and that we can specify the proportion of scores that are above or below a certain value, such as a score of 1.96 in a standardized normal distribution. Statements about likelihoods associated with scores from a normal distribution use a density function for the distribution, which is defined as

$$L_i = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{\frac{-.5(X_i - \mu)^2}{\sigma^2}}$$

where L is the likelihood of score i, X is the variable in question, μ is the mean of the distribution, σ is the standard deviation of the distribution, π is the mathematical constant pi, *e* is the constant associated with the Napierian logarithm, and the likelihood describes the relative height of the normal curve at the value of X. If scores are normally distributed with a mean of 100 and a standard deviation of 10, then, using the above formula, I find that the likelihood of a score of 99 is 0.0397; for a score of 90, it is

0.0242.[3]

When we calculate the mean and standard deviation of a sample to estimate the population mean and population standard deviation of a variable, there are well known formulas we use. Maximum likelihood uses a different approach than invoking these formulas when estimating such population parameters. In maximum likelihood estimation, we seek to identify the population parameter values that have the highest probability of producing a particular sample of data, namely the data we have collected. Suppose I formulate an *a priori* "model" that states scores for a variable in a population are normally distributed with a mean of 95 and a standard deviation of 15. How can I "test" this model or conjecture? Recall from probability theory that the joint probability for a set of independent events is the product of the individual probabilities. For example, the joint probability that for two flips of a coin I will obtain two heads is (0.50)(0.50) = 0.25. This same principle holds for likelihood values so that the overall likelihood for the above "model" producing the sample data is simply the product of the individual likelihoods associated with the respective sample scores. Stated mathematically, it is

$$\mathbf{L} = \prod L_i$$

where **L** (bold faced) is the model likelihood, L_i is the likelihood of a given score *i* in the target distribution, and $\prod$ is the multiplication operator (like a summation operator, $\sum$).

To calculate the likelihood of a model that posits the sample data come from a normal distribution with a population mean of 95 and a standard deviation of 15, I can use the values 95 and 15 in the above density function to calculate the likelihood of each observed data value in the sample data. I then multiply each likelihood by one another to yield **L**. The resulting value of **L**, the model likelihood, is typically a very small number that is difficult to work with because of rounding error, even with access to powerful computers. Because of this, it is common to use a more workable metric by calculating the natural log of the individual likelihoods and then summing these logged values into an overall index of the likelihood of the sample data, known as a **log likelihood** (LL). I signify individual level log likelihood values as LL_i and the summed values using a bold faced **LL** with no subscript. **LL** and LL_i quantify relative probability but they do so on a metric that is computationally convenient. The values of **LL** and LL_i are always negative because the natural log of numbers less than 1.00 (which L values almost always are) are negative. For example, the natural log of 1.00 is zero, the natural log of 0.50 is -0.69, the

[3] Technically, a likelihood as described above is not the same as a probability; it describes the height of the normal curve at a particular score value and represents the relative probability of obtaining a given score from a normally distributed population with a particular mean and variance.

natural log of 0.25 is -1.39, and the natural log of .01 is -4.61. The closer a **LL** is to zero (i.e., the less negative it is), the higher its relative probability.

In maximum likelihood estimation, we calculate the **LL** for different "models," in this case models that posit different values of the population mean and standard deviation. Maximum likelihood then choses the values for the means and standard deviations as population estimates for the model with the highest **LL**. The "trial and error" search process for the model with the highest **LL** is iterative but it is not haphazard; rather, it is a sophisticated algorithm that quickly and efficiently zeros in on the population estimates that have the highest **LL**.

Although this explanation of maximum likelihood estimation glosses over many technical details, it conveys the general idea of comparing log likelihoods for multiple models and then choosing parameter estimates for the model with the highest **LL**, i.e., that have the highest likelihood of having produced the sample data.

Full Information Maximum Likelihood for Missing Data

I now use the above ideas to explicate the FIML approach to missing data analysis. Suppose a researcher plans to fit a single factor model to 6 indicators, X1 to X6, and that each indicator ranges from -3 to +3. The researcher believes a single underlying latent variable will fully account for the population covariance matrix and population means for the observed variables. Consistent with maximum likelihood analysis, Mplus evaluates different one-factor models that vary in their values of the parameters of the factor model (e.g., the values of their factor loadings). The idea is to find the model with the values that best reproduce the observed item means, variances, and covariances, i.e., the model that has the highest log likelihood of having produced the sample data. As part of this search process, we make the assumption that the six variables are multivariately normally distributed in the population so that we can calculate the relevant log likelihoods of the different models, just as I did above. However, now instead of a density function for a normal distribution, I must use the density function for a multivariate normal distribution because I have 6 variables not one. For a given model, I calculate the log likelihood for each case's data assuming the model is correct and then I cumulate the individual log likelihoods into an overall log likelihood for the model. The formula for the individual log likelihoods, LL_i, for a multivariate normal distribution is (Enders, 2010):

$$LL_i = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln(|\sum|) - \frac{1}{2}(X_i - \mu)^T \sum^{-1}(X_i - \mu) \qquad [26.1]$$

where k is the number of variables, $\sum$ is the (hypothesized) population covariance matrix between the k variables based on the sample data, X is a k by1 vector of the X scores for

individual *i*, $|\sum|$ is the determinant of the covariance matrix, μ is a k by 1 vector of (hypothesized) variable means, and T is the matrix transpose operation. Although I make use of this equation here, you need not concern yourself with the details of it because Mplus does all the relevant calculations for FIML for you.

Now, suppose that X1 to X6 have missing data. How do I deal with this when calculating the individual log likelihoods? Suppose for the first iteration of the search process, Mplus evaluates a model that has the following covariance matrix, $\sum$:

	X1	X2	X3	X4	X5	X6
X1	1.10	0.71	0.73	0.33	0.32	0.31
X2	0.71	1.40	0.72	0.36	0.35	0.34
X3	0.73	0.72	1.30	0.39	0.38	0.37
X4	0.33	0.36	0.39	1.20	0.74	0.76
X5	0.32	0.35	0.38	0.74	1.60	0.75
X6	0.31	0.34	0.37	0.76	0.75	1.50

with mean estimates of 0.06, 0.03, 0.02, 0.01, 0.04 and 0.05 for μ for X1 through X6, respectively. I need to calculate the value of **LL** for this particular model. Here are the scores for the first five individuals in the data set on X1 through X6:

Case	X1	X2	X3	X4	X5	X6
1	-1	-2	-2	1	2	2
2	-1	-2	-1	0	0	.
3	2	.	.	0	2	0
4	2	0	1	.	.	2
5	-1	0	.	.	.	.

Individual 1 has complete data for all 6 variables. Individual 2 has missing data on X6 as indicated by a dot. Individual 3 has missing data on X2 and X3. And so on.

My first step is to calculate the log likelihoods for each individual. Individual 1 has complete data so I apply Equation 26.1 to his or her 6 scores using the model $\sum$ and μ matrices from above. The result for the individual is -10.93. For the second individual, there are only 5 scores, with missing data occurring on X6. I can still calculate a log likelihood for this individual using Equation 26.1, but now I focus on the five variables on which s/he provided data. I set k to 5 in Equation 26.1 and I use this individual's five scores and the original model $\sum$ and μ matrices but with variable X6 removed from them. $\sum$ thus becomes

	X1	X2	X3	X4	X5
X1	1.10	0.71	0.73	0.33	0.32
X2	0.71	1.40	0.72	0.36	0.35
X3	0.73	0.72	1.30	0.39	0.38
X4	0.33	0.36	0.39	1.20	0.74
X5	0.32	0.35	0.38	0.74	1.60

and the mean values, μ, are 0.06, 0.03, 0.02, 0.01, and 0.04 for X1 through X5. I find the log likelihood for this individual to be -6.16. Note that in this case I calculate the log likelihood of the individual's scores in a 5 variable multivariate normal distribution rather than a 6 variable multivariate normal distribution, where the 5 variable system is a subset of the 6 variable system. I repeat this process for each individual, adapting the value of k and the elements of X, $\sum$, and μ to the missing data pattern of each individual. For individual 5, for example, I base the LL_i on a bivariate normal distribution with 2 X scores (-1 and 0), the values 0.06 and 0.03 for μ, and the $\sum$ matrix

	X1	X2
X1	1.10	0.71
X2	0.71	1.40

Here is the table of derived LL_i values for the first five cases:

Case	X1	X2	X3	X4	X5	X6	LL_i
1	-1	-2	-2	1	2	2	-10.93
2	-1	-2	-1	0	0	NA	-6.16
3	2	.	.	0	2	0	-7.24
4	2	0	1	.	.	2	-7.21
5	-1	0	.	.	.	.	-2.59

I sum the log likelihood values across all individuals to obtain **LL** for the model. I then repeat this process at the next "trial and error" iteration in the search process using different model parameter values that yield different values of $\sum$ and μ. Ultimately, across all the iterations, I choose the model with the parameter estimates that have the highest **LL**. Such is FIML.

Mplus by default applies FIML missing data algorithms to endogenous variables in the specified SEM model. It applies listwise deletion to exogenous variables so that its default treatment of missing data is a mixture of listwise and FIML approaches. Mplus

uses this strategy, in part, because of the underlying assumptions of the statistical model and the fact that in classic regression modeling the properties of missing data for predictors often have few implications for estimation and inference (see Rönkkö, 2020). FIML missing data strategies can be applied to exogenous variables if listwise deletion is too harsh in terms of data deletion, but you must then override the Mplus default approach. To do so, add syntax to Mplus code to estimate explicitly the variance of each exogenous variable by specifying the name of the exogenous variable on a command line. Alternatively, you can refer to the exogenous variable in a `WITH` statement with another exogenous variable. For an example, see Chapter 11. Applying FIML to exogenous variables comes at some cost as FIML invokes distributional assumptions for the predictors. FIML tends to be reasonably robust to non-normality but the issue of FIML assumptions extends to predictors if you use FIML on them. To be consistent with the underlying statistical theory, if you apply FIML to one exogenous variable, then you should apply it to all exogenous variables in the model by mentioning the variances of them all or incorporating them all into a `WITH` statement.

Table 26.3 presents results for the key path coefficients of the chronic pain model as applied to the complete case data using the MLR option in Mplus and also when applied to the data with missing values using the FIML default option. The results are close to one another. The estimated standard errors are slightly higher and the critical ratios are slightly lower for the data with missing values as compared to the complete data because of the loss of information that comes with missing data. Nevertheless, the FIML approach appears quite serviceable in this case. Of course, I would not be able to do such a comparison in practice because I would not have complete data available to me. However, the exercise here is revealing.

Table 26.3: Results for Complete Case and FIML Analyses

Parameter	Coefficient	Standard Error	Critical Ratio	p Value
T→RS: Complete data	0.973	0.096	10.099	<.001
T→RS: Missing data FIML	0.949	0.098	9.652	<.001
T→KO: Complete data	0.931	0.099	9.357	<.001
T→KO: Missing data FIML	0.942	0.101	9.356	<.001
T→CR: Complete data	1.091	0.095	11.468	<.001
T→CR: Missing data FIML	1.116	0.097	11.450	<.001
RS→CP: Complete data	-0.900	0.117	7.720	<.001

RS→CP: Missing data FIML	-0.959	0.131	7.316	<.001
KO→CP: Complete data	-1.156	0.126	9.139	<.001
KO→CP: Missing data FIML	-1.151	0.145	7.954	<.001
CR→CP: Complete data	-0.836	0.124	6.745	<.001
CR→CP: Missing data FIML	-0.897	0.135	6.666	<.001
DE of T→CP: Complete data	0.074	0.278	0.268	0.789
DE of T→CP: Missing data FIML	0.053	0.319	0.167	0.867
TE of T→CP: Complete data	-2.789	0.273	10.227	<.001
TE of T→CP: Missing data FIML	-2.942	0.293	10.032	<.001

(notes: T = treatment, RS = relation skills, KO = knowledge of opioid mediation use, CR = cognitive restructuring, CP = chronic pain, DE = direct effect, TE = total effect; coefficients are unstandardized. The population values for the effects of the treatment on each mediator is 1.0 and for the effect of each mediator on CP is -1.0)

FIML and Non-Normality. The FIML missing data approach for SEM generally works well when the assumption of multivariate normality is met and the tested model is correctly specified. For ordinal/binary endogenous variables, FIML approaches to missing data use the latent propensity framework discussed in Chapter 5 in conjunction with a maximum likelihood based probit link.[4] If the data are either MCAR or MAR and the above assumptions are met, then FIML-based coefficient estimates are asymptotically unbiased and efficient when standard errors are based on the observed information matrix.[5] As the variable distributions deviate from multivariate normality or the model is misspecified, parameter and standard error estimation can be undermined. Fortunately, simulation studies suggest that FIML is fairly robust to violations of multivariate normality, especially when used in conjunction with robust estimators, such as the MLR option in Mplus. As examples, Yuan, Yang-Wallentin and Bentler (2012) evaluated the effects of non-normality on parameter and standard error bias for simple mean/covariance models consisting of either two variables or five variables. They found that robust FIML tended to show negligible bias when estimating variable means, variances, and covariances for sample sizes of 100 or larger when the overall proportion of missing data was 0.05, 0.15 or 0.25, even in distributions where one or more of the variables was highly leptokurtic (near 10) coupled with skewness values near 2. Jia and Wu (2019) studied the use of robust FIML for variables measured with five categories for sample sizes of 300 and 600 with overall missing data proportions of 0.15 and 0.30 under MCAR

[4] WLSMV for binary outcomes as implemented in Mplus does not use FIML. It uses pairwise deletion for missing data. It works best when data are MCAR as opposed to MAR (see Asparouhov & Muthén, 2010c).

[5] An alternative strategy is to obtain standard errors from the expected or predicted information matrix, but this generally is not recommended with missing data (Enders, 2010). Mplus uses the observed information matrix by default.

and under different variants of MAR. The variable distributions studied were symmetric, moderately asymmetric and severely asymmetric. The tested SEM model was a three latent variable multiple regression model (one latent outcome and two latent predictors) with three indicators for each latent variable. The robust version of FIML performed well for both MCAR and MAR conditions in term of confidence interval coverage and lack of bias in parameters and standard errors for factor loadings and path coefficients when distributions were symmetrical or moderately asymmetrical. Performance degraded somewhat for severely asymmetric data but it still was not unreasonable. Other studies also have been supportive of robust FIML under conditions of MCAR and MAR with non-normal data (e.g., Li & Lomax, 2017), although method performance becomes more questionable when the MAR missingness occurs mainly in the heavy tail of the distributions and the proportion of missing data is large (30% or greater; see Enders, 2001; Savalei & Falk, 2014).

Zhang and Savalei (2019) found that the use of (traditional, non-MLR) FIML tended to yield biased results for the RMSEA and CFI global fit indices in SEM. The degree of bias was impacted by the location of missing data relative to the location of model misfit, the degree of misfit in the hypothesized model, the proportion of missing data, the number of variables with missing data, the type of missing data mechanism, and the number of missing data patterns. In general, the distortions were most likely to operate when the proportion of missing data was 30% or higher. See also the work of Zhang and Savalei (2022).

FIML and Product Terms. Most Monte Carlo evaluations of FIML have evaluated linear models without product terms. There has been recent interest in the performance of FIML when product terms are part of the model because product terms are non-normally distributed even when their component parts are normally distributed. This non-normality violates FIML assumptions. Enders, Baraldi and Cham (2014) studied the case of two predictors (X and Z) and their interaction (XZ) in the presence of normal and non-normal X and Z distributions (e.g., no non-normality, skewness and kurtosis values of 2 and 6), a missing data rate of 20%, and the case where the predictor X was MCAR or MAR depending on Z. They used traditional maximum likelihood rather than the robust version of FIML. Enders et al. found that when data were MCAR and for both small and large interaction effect sizes, FIML coefficient estimates tended to be free of meaningful bias and produced reasonable 95% confidence interval coverage across all conditions studied. When the data were MAR, FIML also provided good confidence interval coverage for the product term coefficient but its performance for tests of significance for the component coefficients was somewhat uneven when non-normality for the components was concentrated in the upper tails of the distribution.

Zhang and Wan (2015) conducted simulations to evaluate FIML as applied to linear regression with product terms but under a wider range of conditions than Enders et al. (2014). They included two forms of MAR rather than one, explored a wider range of sample sizes, a wider range of interaction effect sizes, their non-normality in the component parts had larger kurtosis, and there were different rates of missingness (15%, 30% and 60%). Like Enders et al., they used traditional maximum likelihood estimation rather than robust maximum likelihood estimation. Zhang and Wan (2015) found the same basic results as Enders et al. for the case of N = 100 or greater when the data were MCAR. They also found that under MAR, FIML performed reasonably well for N>100 when the true interaction effect was nil, i.e., its true coefficient was zero. However, as the interaction coefficient became stronger, the performance of FIML tended to degrade under MAR, with the degradation becoming stronger the more that missingness on X was dependent on Z.

Cham et al. (2017) conducted simulations to evaluate FIML as applied to latent variable interaction analysis with missing data on the interaction indicators. They found that FIML performed well with respect to Type I errors and confidence interval coverage for both the LMS interaction method (the default in Mplus) and the product indicator approaches of Marsh et al. (2004) across a wide range of sample sizes, rates of missing data, interaction effect sizes, and both MCAR and MAR missingness. This was true of both the product term coefficient and the coefficients for the component parts.

For categorical moderators with few values in which interactions are evaluated using multi-group SEM (Chapter 20), the default in Mplus is to apply FIML within each group separately but then to select parameter estimates that maximize the overall multiple group log-likelihood. This missing data approach generally works well when the model is correctly specified and under normality or non-normality with robust FIML (see Enders, 2010, 2022; Enders & Gottschall, 2011).

If one of the variables in a product term is binary and if dummy coding for the binary variable is used (e.g., 0 = control, 1 = treatment), then the product term for individuals in the group scored 0 on the binary variable must equal 0 even if they have missing data on the other variable in the product term. In such cases, you can set the product term to zero for individuals in the group scored 0 on the binary variable.

In sum, although product terms introduce possible assumption violations of multivariate normality, FIML tends to work reasonably well for missing data in such cases as long as the data are MCAR and the total sample size is 100 or greater. If the data are MAR and the interaction effect is strong, then confidence interval coverage can be adversely affected, but statistical power of the coefficient for the product term likely will remain high because of the strong interaction to begin with. FIML is viable for dealing

with missing data when analyzing moderation, although it is not without shortcomings.

FIML and Auxiliary Variables. Sometimes MAR can be non-trivially violated when conditioning on your model variables but it is still possible to achieve the conditional independence conditions of MAR through the use of what are known as auxiliary variables. **Auxiliary variables** usually are not of substantive interest but they are included in an analysis to correct the missing data dependencies that can otherwise undermine statistical inference.[6] For example, when planning your study, you might carefully think about the unmeasured U variables in Figure 26.1e, identify those that might undermine missing data conditional independence, and then measure them for purposes of statistical control to turn a MNAR scenario into a MAR scenario through the use of the auxiliary variable option in Mplus, whici I introduce shortly.

An issue that arises when using auxiliary variables is how to control for them analytically. Auxiliary variables are not part of one's substantive model; they are measured for methodological reasons as part of one's theory of missingness. Enders (2006, 2022) describes a strategy to control for them without altering the substantive model of interest. The strategy is known as the **method of saturated correlates** and is implemented in Mplus through the addition of one command, the `AUXILLARY` command. The command is placed after the `MISSING` command that defines missing data values, as follows:

```
MISSING = ALL (-9999) ;
AUXILIARY = (m) z1 z2 ;
```

The `AUXILIARY` option (coupled with the `m` in parentheses) identifies variables to be used as missing data correlates in addition to the core variables in your model. In the above example, the variables z1 and z2 are treated as auxiliary variables. With the addition of this one command, Mplus applies the saturated correlates method to adjust for the dependencies caused by the auxiliary variables primarily by introducing correlations between the auxiliary variables and other exogenous variables as well as disturbance terms, all in ways that do not change the core structure of the substantive model. Of course, one can formally incorporate auxiliary variables into one's model in more traditional ways as covariates but auxiliary variables typically are seen as missing data-based nuisance variables that are of little substantive interest.

Early advice for auxiliary variables was to adopt an inclusive approach erring on the side of including more variables rather than fewer variables when trying to eliminate dependency controls (Schafer & Graham, 2002). Schafer (1997) notes that reasonable candidates for use as auxiliary variables are ones that (a) are highly correlated with

[6] Auxiliary variables have additional functions in multiple imputation. I discuss these functions below.

missingness on the target variable (i.e., the R_Y indicator) and (b) that also are highly correlated with the target variable Y (e.g., Weight in the sex and Weight* example). Schafer (1997) and Raykov and Marcoulides (2013) suggest a data driven approach for identifying auxiliary variables that sets a correlation threshold for each of the above associations; if a variable exceeds one or both of the thresholds it becomes a viable auxiliary variable candidate. Guidelines for defining threshold values vary. For example, van Burren et al. (2010) suggest ± 0.10 whereas Collins et al. (2001) suggest ±0.40.

Recently, inclusive approaches to auxiliary variables have been questioned by several methodologists. Mustillo (2013) notes that the impact of auxiliary variables is a complex function of the magnitude of the correlation between the auxiliary variables and the variables where missing data are present, the proportion of missing cases and the pattern/type of missingness. In multiple simulations, Mustillo (2013) found that auxiliary variables were primarily beneficial when missing data rates were above 50% and the correlation between the auxiliary variable and Y was quite high (e.g., above 0.90), a combination that she argues rarely occurs in applied research (Hardt, Herke & Leonhart, 2012l von Hippel & Lynch, 2000). This suggests that the use of auxiliary variables may not be all that helpful in realistic empirical settings (see also Mustillo & Kwon, 2015). Thoemmes and Rose (2014) identified scenarios were the addition of auxiliary variables increased rather than reduced bias in parameter estimates in the presence of collider dynamics (per Chapter 2; see also Asendorpf et al., 2012). Thoemmes and Rose caution researchers against atheoretical inclusions of auxiliary variables in their modeling. In Mplus, if you use auxiliary variables via the `AUXILIARY` command, you cannot obtain model modification indices nor can you use bootstrapping, two important tools for RET analysis. My own recommendation is similar to Thoemmes and Rose (2014); think through your missing data theory carefully and resort to auxiliary variables with FIML if you have compelling evidence or suspicions that they do, in fact, matter. Bottom line is that you will encounter conflicting advice on whether to be inclusive or exclusive in your approach to auxiliary variables. My advice: Be wary of atheoretical partialling.

Concluding Comments on FIML. In sum, across a wide range of simulation studies, robust FIML strategies for missing data fare well and seem to be reasonably robust to violations of many forms of multivariate normality. They typically can be used to good effect in SEM modeling of RETs with missing data, although they are not perfect.

Bayesian Full Information Approaches to Missing Data

An alternative to FIML that also uses full information algorithms for missing data is based in Bayesian structural equation modeling (BSEM, see Chapter 8). The approach as implemented in Mplus assumes data are MCAR or MAR and that the data are

multivariately normally distributed. Asymptotically and with non-informative priors, the Bayesian approach produces results that closely approximate FIML but its performance with smaller sample sizes has not been adequately explored.

The statistical underpinnings of missing data BSEM are described in Lee (2007) and Asparouhov and Muthén (2010b). In Bayesian modeling, a major goal is to make probabilistic claims about unknown parameters conditioned on a set of knowns, namely the collected data and the prior distribution. Instead of log likelihoods, Bayesian SEM uses likelihoods tied to integrals in calculus. One posits a prior distribution for the parameters of interest, usually in the form of non-informative priors. Posterior parameter distributions are then derived taking into account both the prior distributions and the collected data vis-a-vis iterative methods, such as a Markov Chain Monte Carlo (MCMC) algorithm (see Chapter 8). Like traditional parameters in BSEM, missing data are treated as "unknowns" which the analyst also specifies prior distributions for. Typically, these prior distributions take the form of non-informative priors but they do not have to. In Bayes modeling, the observed data are known with certainty but the missing data are not.

The underlying mathematics of Bayesian missing data methods are beyond the scope of this introductory book. I refer you to Lee (2007), Asparouhov and Muthén (2010b), Jia (2016) and Muthén, Muthén, and Asparouhov (2016) for details. However, to give you a sense of the underlying mechanics, I provide general characterizations here (I assume you are familiar with the material on BSEM from Chapter 8). Some readers may want to skip the remainder of this paragraph. Let Θ be a vector of model parameters to be estimated. Initial information about each element of Θ is contained in the prior distribution for those elements. MCMC samples parameter values from plausible posterior distributions for the Θ and then evaluates model likelihoods for those sampled values. Different sampling strategies are possible. Mplus uses one known as the **Gibbs sampler** (Geman & Geman, 1984). After a set of initial values are evaluated, the posterior estimates are updated in each successive iteration relative to the prior iteration, all with the goal of increasing model likelihoods. These are referred to as **Monte Carlo iterations** (Gilks et al., 1996). Decisions are required about the number of Markov "chains" to be used during the search process and the number of iterations of the sampler. Each chain samples values from a different location of the plausible posterior distributions. Ideally, analyses using the different chains will converge toward the same posterior mean/median value for each Θ parameter. Once the analysis of chains has stabilized, the iterations prior to the stabilization (called the "burn-in" phase) are discarded. Posterior values for the elements of Θ are then computed using the post burn-in iterations. Determining convergence in MCMC estimation is challenging because it seeks to converge on a distribution rather than to a point estimate. This is why BSEM

requires that we examine diagnostics indicative of convergence (see Chapter 8). With missing data, MCMC sampling imputes the unknown data points with plausible values at a given iteration with the goal of maximizing the overall joint model likelihood. The process makes use of the full information likelihoods from the broader model to inform the imputations. As such, the Bayesian approach to missing data is both FIML-like and multiple imputation-like in character. Mathematically, it operates in the same spirit as FIML but from an estimation standpoint it operates somewhat like imputation.

Full information based BSEM missing data approaches have been evaluated in simulation studies but usually with large sample sizes (500 or greater; see Asparouhov & Muthén, 2010b; Song & Lee, 2002; and Lee & Song, 2004a). The methods generally have fared well with large N. In a study with somewhat smaller N, Jia (2016) applied both robust FIML and BSEM as implemented in Mplus to a three latent variable multiple regression model (one latent outcome and two latent predictors) with three indicators for each latent variable. Non-normality was introduced in the form of three different combinations of univariate skewness and kurtosis: mild non-normality (S = 1.5, K = 3), moderate non-normality (S = 2, K = 7), and severe non-normality (S = 3, K = 21). Sample sizes were either 300 or 600. Missing data were created on two of the indicators for each latent variable. Jia investigated cases where data were MCAR as well as several variants of MAR with the proportion of missing data being either 0.15 or 0.30. In all cases, the robust version of FIML (using MLR in Mplus) performed well in terms of confidence interval coverage and lack of bias in parameters for both factor loadings and path coefficients. BSEM performed well for the cases of mild and moderate non-normality but exhibited relatively poor performance for credibility interval coverage under severe non-normality.

As discussed above, exogenous variables in FIML are listwise deleted by Mplus by default and this also is true for BSEM. Exogenous variables can be formally brought into the model by including syntax that tells Mplus to estimate their variance. In FIML, if this is done, then binary exogenous variables are treated as if they are continuous and normally distributed even though they do not have these properties. In BSEM, it is possible to treat the binary exogenous variables more appropriately using the probit-based latent propensity framework described in Chapter 5. To do so, you specify the binary or categorical (ordinal) exogenous variables on the `CATEGORICAL` subcommand (Mplus does not allow this for traditional FIML) and you also specify to estimate the variance of the variable. For an example, see Chapter 9 in Muthén et al. (2016).

Table 26.4 presents results for the key path coefficients of the chronic pain model as applied to the complete case data using BSEM with non-informative priors and also when applied to the same data where some of the complete case data has been converted to

missing values, as described earlier. The Mplus syntax is identical to that for the FIML analysis in Table 26.2 except the line that specifies the estimator (line 10) is changed to

```
ESTIMATOR = BAYES; BITERATIONS=100000 (50000); BCONVERGENCE =.01;
```

and `SAMP` and `MOD(ALL 4)` are removed from the `OUTPUT` line, per Chapter 8. The `CINTERVAL` option is changed to `CINTERVAL(HPD)`. The complete data results and the missing data results are again quite close to one another except for the slightly higher posterior standard deviation and wider confidence intervals due to the presence of less information for the missing values data set. The general conclusions are the same as those for the FIML analysis. To be sure, the nature of the statistical parameters are somewhat different because we have shifted to a Bayesian framework but the conclusions across analyses are comparable.

Table 26.4: Results for Complete Case and Bayesian Analyses

Parameter	Coefficient	Posterior SD	95% Credible Interval
T→RS: Complete data	0.973	0.096	0.784 to 1.163
T→RS: Missing data Bayes	0.949	0.099	0.758 to 1.145
T→KO: Complete data	0.930	0.103	0.728 to 1.131
T→KO: Missing data Bayes	0.943	0.103	0.740 to 1.145
T→CR: Complete data	1.091	0.099	0.902 to 1.281
T→CR: Missing data Bayes	1.115	0.099	0.919 to 1.309
RS→CP: Complete data	-0.900	0.122	-1.141 to -0.661
RS→CP: Missing data Bayes	-0.959	0.140	-1.233 to -0.684
KO→CP: Complete data	-1.155	0.122	-1.395 to -0.916
KO→CP: Missing data Bayes	-1.147	0.139	-1.418 to -0.874
CR→CP: Complete data	-0.835	0.126	-1.084 to -0.589
CR→CP: Missing data Bayes	-0.895	0.140	-1.171 to -0.624
DE of T→CP: Complete data	0.076	0.305	-0.524 to 0.667
DE of T→CP: Missing data FIML	0.044	0.342	-0.630 to 0.718
TE of T→CP: Complete data	-2.788	0.281	-3.340 to -2.236
TE of T→CP: Missing data FIML	-2.944	0.302	-3.535 to -2.351

(notes: T = treatment, RS = relation skills, KO = knowledge of opioid mediation use, CR = cognitive restructuring, CP = chronic pain, DE = direct effect, TE = total effect; coefficients are unstandardized. The population values for the effects of the treatment on each mediator is 1.0 and for the effect of each mediator on CP is -1.0)

Single Imputation Approaches

Another approach to dealing with missing data is to impute replacement score(s) for the missing score(s) of each individual. With single imputation strategies, after a score is imputed for each missing score, the data are analyzed as if it consisted of complete cases. Most methodologists discourage the use of single imputation methods because they usually lead to underestimation of standard errors. The problem is that they treat the imputed score as if there is no uncertainty attached to it, as if the person or case in question actually provided that score. In reality, there is a degree of uncertainty about what the score would have been had the person actually provided the data point and this uncertainty should be reflected in the standard error of the parameter estimate. Multiple imputation methods described in the next section take the uncertainty into account.

One early approach to single imputation is called **regression imputation**. It replaces missing values with predicted scores derived from complete case regression analysis. Specifically, the subset of cases with complete data is used to estimate a regression equation where a target incomplete variable is treated as the outcome and all other variables chosen by the investigator are used as predictors. For example, if there are four variables, X_1, X_2, X_3, and X_4 in the data and one wants to impute missing values for X_2, one isolates individuals with complete data on all four variables and estimates the regression equation (using sample notation):

$$X_2 = a + b_1 X_1 + b_2 X_3 + b_3 X_4 + e \qquad [26.2]$$

This equation is called the **imputation model**, which is then used to generate predicted scores for the incomplete cases on X_2 assuming those cases have scores on X_1, X_3, and X_4 that then allow us to calculate a predicted X_2 value. For each predicted score, random error is added to it based on a normal distribution with a mean of zero and a variance equal to the residual variance in the complete case regression model. Without this latter step, one would impute the same predicted score for X_2 for all respondents who have the same scores on X_1, X_3, and X_4, which would ignore the fact that there is variability in X_2 scores even when X_1, X_3, and X_4 are held constant. Adding such random error takes into account the variability that occurs for the observed scores at any given predictor profile proportional to the magnitude of the estimated error variance.[7]

Regression imputation produces parameter estimates that are unbiased under both MCAR and MAR scenarios, although the estimation of standard errors is typically biased, per my prior discussion. The performance of regression imputation also can

[7] In practice, the regression equations for imputing scores are formulated in complex ways based on different patterns of multivariate missingness. I do not describe those approaches here. For details, see Enders (2022).

depend on specification error in the imputation model. In regression imputation, imputed values can fall outside the range of plausible values of the outcome and they also can take on fractional values even when the outcome metric is integer, a property some analysts do not like.

Another popular approach to single imputation is called **hot deck imputation**. There are many variants of it, some of which fare better than others. A review of hot deck strategies can be found in Andridge and Little (2010). Hot deck imputation replaces missing values on X for a respondent (called the **recipient**) with values on X from a "similar" respondent in the data set (called the **donor**) based on comparability of observed values for the donor and recipient on other variables. I refer to these "other variables" as **matching variables**. In **random hot deck imputation**, the donor is selected randomly from the set of all possible donors in the sample who have identical scores on the matching variables. The matching variables are chosen *a priori*.

An issue with hot deck imputation is the over-use of a single donor if donors can be re-used. This is especially problematic for small sample sizes. Some hot deck methods limit the number of times a donor can be used. As a general rule, hot deck methods tend to break down when the sample size is small and over-use of donors is one reason why.

When matching variables are continuous, it can be difficult to find donors with exact matches because, technically, there are an infinite number of values for continuous variables. Many (but not all) hot deck methods deal with this problem by grouping scores on a continuous variable into categories (usually 10 to 15 categories) to identify donors. In **nearest neighbor hot deck imputation**, a single donor is identified using a mathematical metric of similarity on the matching variables, allowing imputation even if there is not an exact match on the matching variables between donor and recipient.

Hot deck methods have the desirable property that the distribution of imputed X values tends to match the distribution of the observed X values, even if the distributions are non-normal. Nor does it matter if the matching variables are linearly or non-linearly related to the target variable. Unlike regression imputation, hot deck methods do not rely on model fitting to produce imputed scores on the target variable (such as Equation 26.2) nor is random noise arbitrarily added to donated scores. Hot deck imputations are thus less sensitive to misspecification error in the imputation model. In addition, only plausible values can be imputed because values come directly from donors.

Another form of hot deck imputation uses a **random recursive partitioning** (RRP) imputation strategy (Iacus & Porro 2007, 2009). This method identifies donors by analyzing the spatial distribution of observations on a set of matching variables, identifying potential donors who lie in the same multivariate space as the recipient and then imputing a score from a subset of the pool of donors. It uses the logic of regression

trees (Breiman, Friedman, Olshen & Stone, 1984) and focuses on the ordering of values on the matching variables rather than the values per se. Like most other hot deck methods, RRP creates categories for continuous variables and can be used for both numeric and categorical data. It imputes the average value of the *k* nearest neighbors (or, in the case of categorical variables, the mode of the *k* nearest neighbors), where the value of *k* is specified by the analyst. It makes no assumptions about data structure. For details, see Iacus and Porro (2007, 2009).

The major drawback with the hot deck methods, as with most single imputation methods, is that standard errors based on them can be biased, sometimes trivially but sometimes substantially. There are circumstances where single imputations work fine (e.g., when missing data are few and the N is large), but care must be taken in their use.

Multiple Imputation Approaches

Instead of imputing a single score for an individual who has missing data on X, some analysts replace the missing score with several imputed values. The idea is to create different imputed values reflective of the uncertainty associated with the imputation model because, in reality, we can't be certain what the individual's true score is. As an example, I could use the regression imputation approach described earlier to create 5 different data sets, with each data set potentially having different imputed values for a given individual because of the randomly generated error scores. The multiple imputation approach requires that you then analyze each of the five data sets separately and combine the parameter estimates from each of the separate analyses to yield a final parameter estimate (and standard error). Let me provide an example to make this concrete.

Suppose you want to conduct a multiple regression analysis based on the model $Y = a + b_1 X + b_2 Z + e$, where e is a residual term. You are interested in the values of b_1 and b_2 and whether each is statistically significant. There are missing data on one or more of the variables for some of the cases. You use a regression imputation algorithm to impute scores for the missing data for each case. You repeat this process five times to generate five different data sets, D1, D2, D3, D4, and D5. Again, the scores imputed for a given variable for a given individual vary across the data sets because each error score is a different random draw from the distribution of residuals. In each data set, you conduct the above regression analysis and note the values of b_1 and b_2 as well as their standard errors. This might yield the following results from the five regression analyses:

Data Set	b_1	se_1	b_2	se_2
D1	1.3	.22	12.4	1.25
D2	1.8	.56	12.7	1.22
D3	2.2	.33	13.1	1.37
D4	1.1	.37	12.1	1.44
D5	2.9	.42	13.0	1.26

where *se* is the standard error associated with a given coefficient. In the multiple imputation strategy, we pool the information for the b_1 estimates across the five data sets to derive a single estimate of b_1 and we also pool the estimates of the standard errors for b_1 to derive a single estimate of the standard error for b_1. Then, an additional correction term is added to the standard errors to reflect the variability of the regression coefficients that occur across the imputed data sets. The formula for pooling is complex and is described in Enders (2010; see Bartlett, 2021 for a discussion of strengths and weaknesses of the approach). I provide a program on my website called *Combine MI estimates* that applies the approach. The researcher then uses the pooled estimates and standard errors to perform significance tests and calculate confidence intervals.

A variant of multiple imputation regression modeling is called **stochastic Bayesian regression modeling**. In addition to generating scores by incorporating random noise based on the residual term of the imputation model, methods also are used to represent the uncertainty associated with knowledge of the true population regression coefficients of the imputation model. Thus, random disturbances also are introduced into the regression coefficients themselves in the spirit of Bayesian hyperparameters discussed in Chapter 8. The introduction of randomness at this step is called the **P step**. For elaboration, see Enders (2010) and White, Royston, and Wood (2011).

All multiple imputation methods essentially involve (1) specifying an imputation model, (2) generating multiple imputation data sets, (3) analyzing each imputed data set, and (4) pooling results across the imputed data sets. Fortunately, Mplus automates most of these tasks to make multiple imputation analysis straightforward, as I now discuss.

Bayesian Multiple Imputation

Mplus allows you to generate multiple imputation data sets using a Bayesian approach. The method is distinct from the full information Bayesian strategy I described above, as I elaborate shortly. Bayesian multiple imputation is the default method used by Mplus for multiple imputation. You generate imputed data with Mplus using either prespecified models that Mplus calls **unrestricted H1 models** or using an a priori model of your

choice that is amenable to BSEM. Mplus calls the latter models **restricted H0 models** or, more simply, **H0 models**. The default H1 model is called the **covariance model**.

The covariance H1 model uses Bayesian methods to directly model the means, variances, and covariances among your model variables. There are no path coefficients or error variances in the model. The H1 covariance model can readily handle both continuous and categorical (binary or ordinal) variables. The underlying model is general so that model misspecification at the imputation stage cannot occur but it sometimes has a large number of parameters that make convergence difficult. This tends to occur when there are many variables that are a mixture of categorical and continuous variables. One typically uses a restricted H0 imputation model in place of the H1 model if the H1 model fails to converge or if one believes the H0 model is correctly specified and will yield more efficient estimates than the H1 model, a point I elaborate below. The imputation process for both H0 and H1 models, like FIML, assumes the variables are multivariately normally distributed. As such, concerns about non-normality apply to Mplus Bayesian imputation just as they do for FIML. Mplus also offers multilevel counterparts to the H1 model but these models often have convergence issues. Given this, it is much more common to use H0 imputation for multilevel models.

The imputed data for the H1 covariance model are generated after the MCMC sequence of Bayesian estimation has converged. Given convergence, Mplus conducts an additional 100 MCMC iterations and stores the generated imputed values at the last step for the missing data as the "filled in" values for the first imputed data set. The MCMC process then continues for another 100 iterations and the imputed values at that point are used for the second imputed data set. The process continues until the number of imputed data sets requested by the user is generated. Each data set represents independent draws from the missing data posterior distribution. For statistical details as well as several examples, see Asparouhov and Muthén (2021) and the document *Mplus Imputation Methods* on my webpage. Parenthetically, Mplus offers two other possible H1 imputation models besides the covariance model, a sequential regression model and a regression model. For details of these other models, see Muthén, Muthén and Asparouhov (2016).

Once the imputed values have been generated, you analyze them using a model and estimator of your choosing, such as ML, MLR, or WLSMV. It usually is not a good idea to apply BSEM to Bayesian multiply imputed data because doing so creates technical issues for the credibility intervals. If you want to apply BSEM to your data set, use the full information BSEM method described earlier.

An important consideration in the use of Bayesian imputation is the identification status of the imputation model, in this case the H1 covariance model. To be identified, the sample size of observed data for each variable should equal or be greater than the total

number of variables. For the case where there are 100 individuals in a data set and 25 variables are being imputed, if one of the variables has only 20 observations (i.e., there are 80 individuals with missing data on that variable), then 20 < 25 and the model will be underidentified. One can either remove the offending variable from the imputation model or impute scores for it using a smaller imputation model that has fewer than 20 variables.

For the chronic pain data, I used the Mplus syntax in Table 26.5 to generate 25 imputed data sets using the H1 covariance approach.

Table 26.5: Mplus Syntax to Generate Imputed Data Sets

```
1. TITLE: GENERATE IMPUTED DATA
2. DATA:   FILE = missingexample2a.dat ;
3. VARIABLE:
4. NAMES = ID RS KO CR CP BRS BKO BCR BCP T BSEX ;
5. USEVARIABLES RS KO CR CP BRS BKO BCR BCP T BSEX ;
6. MISSING ARE ALL (-9999) ;
7. AUXILIARY = ID ;
8. DATA IMPUTATION:
9. IMPUTE = RS-BCP T(c) BSEX(c) ;
10. NDATASETS = 25 ;
11. SAVE = missimp*.dat;
12. ANALYSIS: TYPE = BASIC;
```

Lines 1 to 6 are standard Mplus syntax. Line 5 is the traditional `USEVARIABLES` line that contains the variables for the covariance matrix for the imputation model. If you want to carry over any other variables into the imputation data set generated by Mplus that are not part of the focal covariance matrix, then list them on the `AUXILIARY` line, per Line 7. These are not auxiliary variables in the sense I described earlier for FIML. Rather, I am using the auxiliary command merely as way of telling Mplus what other variables to include in the data sets it generates. In the current case, I carry over the variable representing the respondent ID number, `ID`. Line 8 tells Mplus I want to generate imputed data sets. Line 9 tells Mplus what variables from the `USEVARIABLES` line to use in the imputation model. In the current example, I use Mplus shorthand to indicate I want to use all variables on the `USEVARIABLES` list starting with the variable named `RS`, ending with the variable `BCP` and every variable listed between them. I exclude `T` and `BSEX` because they are binary categorical variables and I must identify them as such using the `(c)` notation shown. Line 10 tells Mplus the number of imputed data sets I want to generate, in this case 25. Line 11 tells Mplus where to save the data sets. In this instance, it will be the same folder where the input syntax is stored because I do not specify a folder path. Each data set will be named “missimp” (you can use any name you want) followed by a

number from 1 to 25 (the number of requested data sets) and the tag will be dat (you can use any tag designation you want, such as dat, txt). On Line 12 the analysis type is specified as `BASIC` and the `OUTPUT` line is unnecessary given `BASIC` defaults.

Mplus generates the 25 data sets based on the above syntax. It also generates a file called missimplist.dat, which is the name you gave to each data set followed by the word "list.dat" instead of a number with the tag dat. This file contains the list of names of all the generated data sets in a single column, like this (again, as long as the data files are in the same folder that your Mplus syntax resides, there is no need to specify a folder path):

missimp1.dat
missimp2.dat
missimp3.dat
missimp4.dat
.
.
.
missimp22.dat
missimp23.dat
missimp24.dat
missimp25.dat

As you will see shortly, I use this file to conduct analyses of the chronic pain model data.

It is interesting to compare the actual values that people had in the complete data set with the values that were imputed by Mplus. Table 26.6 presents the actual and imputed values for the first two imputed data sets for the 9 individuals who had missing values on the relaxation skills mediator as well as the 4 individuals who had missing values on biological sex. Some of the imputed values were reasonably close to the actual values and others were "off" by more than we might like. I discuss below ways of increasing the correspondence between the scores.[8] Note that although the actual scores on the variables were integers, the imputed values for the continuous measures have decimals. I discuss below the issue of whether such imputed values should be rounded (in general, they should not be). The categorical/binary variable does not have decimals because I defined it as categorical in the Mplus syntax; the imputed value is based on the continuous latent propensity framework with threshold values per Chapter 5, which taken together, will yield imputed scores that are integer in form.

[8] Interestingly, imputation methods that produce the closest correspondence between the actual and imputed scores do not necessarily separate valid from invalid imputation methods. See van Buuren (2018, section 2.6) for details.

Table 26.6: Actual and Imputed Values for 13 Cases

ID-Variable	Actual Score	Imputed Score 1	Imputed Score 2
83-RS	5	5.249	5.059
88-RS	5	5.372	4.818
98-RS	4	6.937	3.750
136-RS	4	4.724	5.512
137-RS	6	7.381	5.893
138-RS	5	4.059	5.459
221-RS	4	5.082	5.296
249-RS	4	4.309	5.247
269-RS	4	5.571	5.949
1-BSex	1	1	0
2-BSex	1	0	1
84-BSex	0	0	1
165-BSex	1	0	0

After generating the imputed data sets, I need to analyze each one separately using the SEM model of interest and then combine the results to yield my best estimates of each model parameter, their standard errors, confidence intervals, and p values. Fortunately, Mplus automates the process. Table 26.7 presents the relevant Mplus syntax.

Table 26.7: Analysis of Imputed Data Sets

```
1. TITLE: MULTIPLE IMPUTATION MAIN ANALYSIS  ;
2. DATA: FILE IS missimplist.dat ;
3. TYPE = IMPUTATION;
4. VARIABLE:
5. NAMES = RS KO CR CP BRS BKO BCR BCP T BSEX ID ;
6. USE VARIABLES = RS KO CR CP BRS BKO BCR BCP T BSEX ;
7. ANALYSIS:
8. ESTIMATOR = MLR ;
9. MODEL:
10. RS ON BRS T ;
11. KO ON BKO T ;
12. CR ON BCR T ;
13. CP ON BSEX BCP RS KO CR T ;
14. MODEL INDIRECT:
15. CP IND T ;
16. OUTPUT:
17. SAMP STANDARDIZED(STDYX) RESIDUAL CINTERVAL TECH4 ;
```

Line 2 identifies the file that lists the imputed data sets as the input file; Mplus then uses the 25 data sets listed in that file. Line 3 informs Mplus that the data input is in the form of imputed data sets. A line that identifies missing values is not needed because there are none, unless you carried over a non-analytic variable that has missing data. The output line does not include `MOD(ALL 4)` because modification indices cannot be computed for multiple imputation analyses.

There are unique features to the Mplus output when multiple imputation is used that I now discuss. Each global fit index is summarized across the 25 imputed data sets vis-à-vis its mean value and standard deviation. For example, here is the summary output for the chi square fit index:

```
Chi-Square Test of Model Fit

        Degrees of freedom                     18

        Mean                                21.184
        Std Dev                              2.308
        Number of successful computations       25
```

The mean chi square value across the 25 data sets was 21.184 with a standard deviation of 2.308. The `Number of successful computations` was 25, indicating there was model convergence in all 25 analyses. You can't use the mean chi square and degrees of freedom (in this case df = 18) to calculate a valid p value for the statistic (see Asparouhov & Muthén, 2010d, for why this is the case); attempts to evolve a formal global chi square test in multiple imputation modeling are controversial. In theory, one hopes the reported chi square values in the imputed data sets will each be close to or less than the degrees of freedom, but this is only a crude indicator of model fit.

Here is the summary information for the RMSEA:

```
RMSEA (Root Mean Square Error Of Approximation)

        Mean                                 0.022
        Std Dev                              0.011
```

The average RMSEA was 0.022 in the 25 data sets with a standard deviation of 0.011. Mplus also provides information about the distribution of the fit index across the imputed data sets in the form of a cumulative distribution that reflects the cumulative proportion of RMSEA values that were less than or equal to selected values. Actually, Mplus does this for each fit statistic. Here is the distribution for the RMSEA:

```
          Value    Function Value

          0.990         1.000
          0.980         1.000
          0.950         1.000
          0.900         1.000
          0.800         1.000
          0.700         1.000
          0.500         1.000
          0.300         1.000
          0.200         1.000
          0.100         1.000
          0.050         1.000
          0.020         0.360
          0.010         0.160
```

The proportion of the 25 RMSEAs that equaled 0.05 or less was 1.00; the proportion of the RMSEAs that equaled 0.02 or less was 0.36; the proportion of RMSEAs that equaled 0.01 or less was 0.16.

The combined path coefficients and their combined multiple-imputed standard errors, critical ratios, and p values are reported per the customary Mplus output format. For example, here is the output for the linear equation predicting chronic pain from the key mediators, the baseline covariates, and the direct effect of the treatment condition:

```
MODEL RESULTS
                                                    Two-Tailed   Rate of
                    Estimate       S.E.     Est./S.E. P-Value    Missing

CP       ON
    BSEX               0.811      0.256      3.172      0.002      0.207
    BCP                0.316      0.173      1.830      0.067      0.160
    RS                -0.960      0.126     -7.617      0.000      0.163
    KO                -1.182      0.145     -8.162      0.000      0.193
    CR                -0.903      0.134     -6.747      0.000      0.159
    T                  0.117      0.327      0.359      0.720      0.194
```

The only difference relative to the traditional Mplus output is the last column labeled "rate of missing." Some Mplus users mistakenly interpret this as the overall proportion of missing data associated with the estimated parameter. It is not this. Rather, it is a technical index that can be used to evaluate if the number of imputed data sets you used is satisfactory. I discuss it in more depth later.

Table 26.8 presents results for the key path coefficients of the chronic pain model as applied to the complete case data using the MLR option in Mplus and also when applied to the Bayesian imputed data that are then analyzed via Mplus with robust maximum likelihood (MLR). The results are similar to one another. As before, the estimated

standard errors are slightly higher and the critical ratios are slightly lower for the data with missing values because of the loss of information that comes with missing data. The results also are comparable to those observed for the FIML approach reported earlier.

Table 26.8: Results for Complete Case and Bayes Multiple Imputation Analyses

Parameter	Coefficient	Standard Error	Critical Ratio	p Value
T→RS: Complete data	0.973	0.096	10.099	<.001
T→RS: Missing data MI	0.946	0.098	9.654	<.001
T→KO: Complete data	0.931	0.099	9.357	<.001
T→KO: Missing data MI	0.931	0.100	9.318	<.001
T→CR: Complete data	1.091	0.095	11.468	<.001
T→CR: Missing data MI	1.117	0.096	11.631	<.001
RS→CP: Complete data	-0.900	0.117	7.720	<.001
RS→CP: Missing data MI	-0.960	0.126	7.617	<.001
KO→CP: Complete data	-1.156	0.126	9.139	<.001
KO→CP: Missing data MI	-1.182	0.145	8.612	<.001
CR→CP: Complete data	-0.836	0.124	6.745	<.001
CR→CP: Missing data MI	-0.903	0.134	6.747	<.001
DE of T→CP: Complete data	0.074	0.278	0.268	0.789
DE of T→CP: Missing data MI	0.117	0.327	0.359	0.867
TE of T→CP: Complete data	-2.789	0.273	10.227	<.001
TE of T→CP: Missing data MI	-2.899	0.298	9.715	<.001

(notes: MI = multiple imputation, T = treatment, RS = relation skills, KO = knowledge of opioid mediation use, CR = cognitive restructuring, CP = chronic pain, DE = direct effect, TE = total effect; coefficients are unstandardized. The population values for the effects of the treatment on each mediator is 1.0 and for the effect of each mediator on CP is -1.0)

Parenthetically, it is possible to combine the generation of imputed scores and model analysis into a single Mplus run, although I prefer to execute them separately. One reason to conduct the analyses separately is because when performed together, Mplus does not allow you to include the `MODEL INDIRECT` command to compute omnibus mediation effects and the total effect. I personally downplay omnibus mediation effects in favor of individual link analysis per my discussions in Chapters 9 through 11. However, if you are interested in omnibus tests (including the total effect, which I usually *am* interested in), then with multiple imputation performed in a single run, you must

explicitly tell Mplus to compute them using the MODEL CONSTRAINT command by specifying the relevant multiplicative functions between coefficients. Here is the combined code for the chronic pain model that uses the MODEL CONSTRAINT command:

```
1. TITLE: ANALYSIS OF CHRONIC PAIN
2. DATA:   FILE = missingexample2a.dat ;
3. VARIABLE:
4. NAMES = ID RS KO CR CP BRS BKO BCR BCP T BSEX ;
5. USEVARIABLES = RS KO CR CP BRS BKO BCR BCP T BSEX ;
6. MISSING ARE ALL (-9999) ;
7. DATA IMPUTATION:
8.    IMPUTE = RS KO CR CP BRS BKO BCR BCP T BSEX(c) ;
9.    NDATASETS = 25;
10. ANALYSIS:
11. ESTIMATOR = MLR ;
12. MODEL:
13. RS ON BRS T (cov1 a1) ;
14. KO ON BKO T (cov2 a2) ;
15. CR ON BCR T (cov3 a3) ;
16. CP ON BSEX BCP RS KO CR T (cov4 cov5 b1 b2 b3 b4);
17. MODEL CONSTRAINT: !calculate omnibus effects
18. NEW(OMRS, OMKO, OMCR, TE) ;
19. OMRS = a1*b1 ;
20. OMKO = a2*b2 ;
21. OMCR = a3*b3 ;
22. TE = OMRS + OMKO + OMCR + b4 ;  ! calculate total effect
23. OUTPUT:
24. SAMP STANDARDIZED(STDYX) RESIDUAL CINTERVAL TECH4 ;
```

Lines 1 to 6 are standard Mplus syntax. Lines 7 to 9 generate the imputed data (note that the data sets are not saved). Lines 10 to 24 execute the model analysis, including the MODEL CONSTRAINT lines to calculate the omnibus parameter estimates.

In sum, Bayesian based multiple imputation strategies are easily implemented in Mplus. At the simplest level, one uses the default H1 covariance model coupled with the BASIC option in Mplus (see Table 26.5) to generate the imputed data sets. You then invoke the automated analysis-then-combine Mplus syntax (see Table 26.7) to execute the analysis. One can use fancier H0 modeling strategies and conduct analyses in a single integrated run, but the simple two step strategy described here is often serviceable.

Chained Equations and Predicted Mean Matching

Yet another approach to multiple imputation is called **chained equations** coupled with **predictive mean matching** (PMM). The chained equation method specifies a separate imputation model for each variable in the target data. As such, it is distinct from the

multiple imputation Bayesian approach of Mplus. For example, I might use linear regression to impute values into X_1, logistic regression to impute scores into X_2 if X_2 is binary, a proportional odds ordinal regression model to impute scores into X_3 if X_3 is ordinal, and so on. In the chained equations approach, I can choose whatever predictors in the data set I want for a given prediction equation. The use of different imputation models for each variable stands in contrast to many traditional multiple imputation methods that use the same imputation model for each variable, namely linear regression in which the target variable is regressed onto all other variables in the model. There are multiple variants of the chained equation approach but Allison (2015) captures the basic ideas vis-à-vis the following steps, characterized here assuming a continuous X variable is being predicted from several continuous Z in a linear equation:

Step 1: For cases with no missing data, estimate a linear regression of X on Z, yielding a set of coefficients b.

Step 2: Make a random draw from the posterior distribution of b based on Bayesian logic to produce a new set of coefficients called b*. This typically takes the form of a random draw from a multivariate normal distribution of the b coefficients and is analogous to the P step described earlier.

Step 3: Using b*, generate predicted values for X for all cases including those with data missing on X and those with data present.

Step 4: For each case with missing X, identify a set of cases with observed X whose predicted values on X are close to the predicted value for the case with missing data.

Step 5: From among those close cases or "potential donors", randomly choose one and assign its observed value to substitute for the missing value on X for the case in question.

Step 6: Repeat these steps as needed to generate the number of desired imputed data sets.

There are technicalities within several of these steps and software differs in how they approach them, but the above captures the spirit of the approach. For details of the PMM version emphasized in this chapter, see van Buuren (2018), van Buuren and Groothuis-Oudshoorn (2011) and Gaffert, Meinfelder, and Bosch (2015). For purposes of this book, I use the R package called *mice* (multivariate imputation by chained equations) to implement the method. I provide a program on my website to produce output that is compatible with required Mplus input for combining results across imputed data sets.

A key issue when applying PMM is specifying the size of the donor pool to use,

which usually is fixed a priori by the analyst. The default in the mice package is five donors, which means that each case with missing data on a given target variable is matched to the 5 cases that have the closest predicted values to the individual with missing data. One of the 5 donors is chosen at random and his or her observed value on the target variable is assigned to the individual with missing data. If you set the size of the donor pool too low, then you lose an element of randomness/uncertainty that should be part of the analysis; think of the extreme case where the donor pool is set to a size of one such that there is no random selection at this step. With a small donor pool, you also run the risk of the same person being a donor multiple times across individuals thereby creating unwanted dependencies in the data. On the other hand, if you set the size of the donor pool too high, then you may impute values from cases that are poor matches to the target case. In general, the performance of PMM tends to degrade when the size of the donor pool is small and there are many tied scores for the predictors among the individuals with missing data. Simulations by Schenker and Taylor (1996) found that fixed donor pools of size 3 performed well whereas Morris et al. (2014) recommended a pool size of 10 for a variety of situations. With large samples, a pool size of 10 probably is useful, but with smaller samples, a value closer to 5 seems better (Allison, 2015).

Gaffert, Koller-Meinfelder and Bosch (2016) developed an adaptive rather than fixed method for specifying the donor pool size, a strategy that they refer to as **midastouch**. It assigns weights to each individual with complete data in the sample based on the distance between the individual and recipient cases predicted values. The weights impact the probability an individual will be randomly selected as a donor. Because all observed cases can be donors, there is no need to specify an a priori donor pool size; the weights do the necessary work. See Gaffert et al (2016) for details of the statistical algorithm. Simulations by Gaffert et al. (2016) suggest that the adaptive method is superior to the fixed donor pool size method when the overall sample size is small. The program on my website includes the midastouch option.

A concern with the chained equation approach is that the conditional distributions for different variables may be incompatible. For example, suppose one uses a proportional odds ordinal regression model for imputing X_1 scores and a linear regression model for imputing X_2 scores. These models are incompatible in the sense that there is no joint distribution for the two variables that satisfactorily capture both conditional distributions. Simulation studies suggest this issue has little consequence in practice (White, Royston and Wood, 2011; van Buuren & Groothuis-Oudshoorn, 2011) but it nevertheless is a concern. Several simulation studies indicate that chained equation approaches fare better than many of the more traditional joint modeling multiple imputation approaches (e.g., Faris et al., 2002; Marshall, Altman, Royston & Holder,

2010). Studies also suggest that PMM tends to be less sensitive than many multiple imputation methods to model misspecification in the form of non-linear relationships and non-normality (Schenker & Taylor, 1996; Morris et al., 2014). However, the approach can be sensitive to extreme skew. Using PMM for censored or truncated data can be problematic. Again, the approach works best with larger sample sizes.

When a variable is ordinal in character, many researchers invoke proportional odds ordinal regression to make the PMM imputations (see Chapter 5). I prefer to use multinomial regression in such cases because ordinal regression is a special case of multinomial regression but with more stringent assumptions, as I discuss in Chapter 5.

In sum, chained equations with predictive mean matching is a popular multiple imputation method and it often can be effectively used to analyze RET data when FIML is not viable. I spare you details here, but when I applied the approach to the chronic pain model using the program on my website in conjunction with Mplus, the results were quite similar to those reported for FIML.

On my webpage, I offer point and click access to the *mice* package that offers PMM; see the program called *Imputation: Chained equations*. The interface uses mice defaults applied to the list of variables that you specify. MICE regresses each variable onto all other variables in the list and then applies linear, logistic or multinomial regression. You can override the defaults and specify unique equations for each variable. For details on how to do this, see the document on the resources tab of my webpage in the document describing MICE and BLIMP software.

Factored Regression and BLIMP

Yet another approach to multiple imputation is based on a synthesis of Bayesian and factored regression logic as articulated by Enders (2022; Keller & Enders, 2021). The approach is implemented in software called BLIMP (Bayesian latent imputation). BLIMP parallels the *mice* program for chained equations but it has the capability of working with latent variables, it uses latent response frameworks to treat categorical variables, and it uses latent group means for multilevel data structures. Unlike the chained equations approach, it does not rely on predictive mean matching nor donor pools but, in spirt, it operates somewhat like chained equations.

The BLIMP program makes imputations analogous to the H0 Bayes approach in Mplus. Like Mplus, the idea is to create imputations that are tailored to a particular analysis model rather than using unstructured models per H1 imputation in Mplus. The difference between Mplus H0 imputation and BLIMP is in the distributional specification of the imputation model. Mplus (typically) invokes a joint multivariate distribution (usually multivariate normal) to make imputations whereas BLIMP works with univariate

distributions using a framework known as factored regressions.

Factored regression can be thought of as using several regression models to characterize the joint distribution of variables. Suppose the joint multivariate distribution between three variables is symbolized as $f(X,Y,Z)$. It turns out that I can express this function as the product of three univariate distributions each corresponding to a regression equation as follows:

$$f(X,Y,Z) = f(Y|X,Z) * f(X|Z) * f(Z) \quad [26.3]$$

The first part of the right hand side of this equation, $f(Y|X,Z)$, is the conditional distribution of Y given X and Z. Stated another way and a little less formally, the term represents a regression model where I predict Y from X and Z. Note that the form of this regression is unspecified; it could be a linear regression, a logistic regression, a probit regression, or some other regression type. The idea behind factored regression is to break up or "factor" a complex joint distribution of variables into the product of less complicated conditional distributions and then specify the form of each function (linear, logit, probit) between the outcome variable on the left of the bar in Equation 26.3 and the predictors to the right of the bar. By decomposing the joint distribution into more manageable subparts, BLIMP gains flexibility during the modeling process. By contrast, joint distribution approaches require that multivariate normality (or some other tractable multivariate distribution) holds across the variables. The joint distribution approach works directly with the term on the left side of Equation 26.3 rather than the component parts on the right side of the Equation. When Mplus invokes Bayesian imputation, for example, it usually does so with reference to the multivariate distribution demarcated on the left side of Equation 26.3.

BLIMP, by contrast, works with the models on the right side of the equation. The advantage (and some would say disadvantage) of doing so is that the distributions on the right side of Equation 26.3 can contain features that are at odds with the multivariate normal distribution. For example, they can contain mixtures of categorical and continuous variables or they can include product terms that violate multivariate normal properties. Both the joint multivariate distribution approach and the factored regression approach rely on robustness to these types of violations when anomalies are encountered, but they do so in different ways. A nice feature of the BLIMP software is that it automates for you much of the required mathematics "under the hood," making the factored regression framework relatively easy to apply.

I generated 25 imputed data sets using BLIMP and input them into Mplus to analyze the chronic pain model per the syntax in Table 26.7. The results were similar to those for the FIML analyses. For more details about BLIMP, see the Resources tab on my

webpage and the BLIMP homepage at https://www.appliedmissingdata.com/blimp.

Robust Multiple Imputation

As discussed in Chapters 5 and 6, traditional linear regression is not outlier/leverage resistant. Outliers with large leverages can distort regression coefficients and, in turn, create implausible predicted values thereby wreaking havoc with imputed values in multiple imputation strategies. Imputed values in such contexts can themselves become outliers that inflate standard errors and bias coefficients in the analysis model. Templ, Kowarik and Filzmoser (2011) have developed a multiple imputation method that makes use of robust regression in place of OLS regression. It can accommodate mixtures of binary, ordinal, nominal and continuous variables and, like chained equations, selects an appropriate robust regression method based on the outcome variable metric. It uses MM regression for continuous outcomes and robust variants of logistic, multinomial, and ordinal regression for other types of variables depending on their metric properties. The approach is described in detail by Templ et al. (2011; see also Kowarik & Templ, 2016). I provide a program called *Imputation: Robust* for it on my webpage. Simulations reported by Templ et al. (2011) attest to its viability. When I applied it to the chronic pain data, it yielded results comparable to FIML because the example does not have outliers.

Special Issues in Multiple Imputation Analysis

In this section, I briefly discuss six issues that often arise when using multiple imputation strategies. These include (a) the number of imputations one should use, (b) rounding, (c) clustered and multi-level data, (d) the use of product terms, (e) additional variables and informative priors, and (e) exploratory analyses and single imputation.

Number of Imputations. An important issue when applying multiple imputation strategies is the choice of the number of imputation data sets to generate. In the early days of multiple imputation methodology, the general wisdom was that 4 to 5 imputation data sets were sufficient (Little & Rubin, 1987). Graham et al. (2007) found that larger numbers of imputations tended to produce somewhat more statistical power when seeking to detect small effect sizes. Bodner (2008) also found this to be the case and described guidelines (based on simulations) for determining the number of generated data sets that take into account, among other things, the proportion of missing data. For example, when the proportion of missing data are less than 20%, fewer than 10 imputations often will work fine. A rough rule of thumb is that the number of imputations should equal the overall percent of missing data (rounded), with a minimum of 5 (von Hippel, 2009; see also White, Royston & Wood, 2011).

More recent literature suggests larger numbers of imputations may be necessary,

depending on the complexity of the model and the patterning of data. Current thinking is that it is best to set the number of imputations higher, in the 20 to 100 range (van Buuren, 2018) especially in low power situations. A key concept in this literature is the **fraction of missing information** (FMI). The FMI is *not* the fraction of values that are missing. Rather, it is a technical concept that is an estimate of the fraction by which the derived squared standard error for the parameter in question computed based on the missing data in the data set would shrink if the squared standard error was computed instead with complete data. It ranges from 0 to 1.00. If it equals 0.10, the missing data based squared standard error would reduce by 10% if it was instead computed using complete data. The smaller the value of FMI, the better because it reflects how much missingness inflates standard errors. Mplus reports the FMI on its output for each parameter in your model.

A useful index that is derived from the FMI is called the **proportional increase in the standard error** (PISE; Enders, 2010). The PISE is an estimate of the proportion by which the standard error for the statistic in question is inflated relative to its hypothetical minimum. More technically, it is a multiplicative factor by which the standard error is multiplied to reflect the inflation caused by using the number of imputation data sets you chose to use. If you use 5 imputed data sets and the PISE equals 1.01, this means that with 5 imputed data sets, the standard error for the statistic in question will be about (1.01 – 1) times 100 or 1% larger than what it would be relative to using an extremely large number of imputations, say a million. If the PISE is 1.05, then the standard error is inflated by 5%. If it is 1.13, then the standard error is inflated by 13%. The closer the PISE value is to 1.00, the better. If it is reasonably close to 1.00, then you have probably used a sufficient number of imputation data sets for minimizing the inflation of the standard error. In the programs on multiple imputation on my website, I provide a program to calculate the PISE from the FMI reported on Mplus output to help you evaluate the number of imputed data sets you have chosen.

von Hippel (2018; see also Allison, 2019) has developed another index to help evaluate the number of imputations to use. The index is based on the replicability of the standard error as a function of the number of imputation data sets. It takes the form of a coefficient of variation and is the percentage by which you would be willing to see the standard error for the parameter change (on average) if the data were imputed again for the same number of imputations you are using. A value of 3.5% means you are willing to tolerate a 3.5% change in the value of the standard error if the data were imputed again using the same number of imputation data sets. (Remember, the standard error changes because of the random processes built into the imputation procedure). In general, the smaller the value of the percent change from one set of imputations to another, the better.

For the chronic pain model, I scanned the “Rate of Missing” values reported on the

Mplus output (which are FMIs) for the Bayesian imputation method and the largest one was 0.20. I used the program on my website called *Mplus imputation analysis* to calculate the PSIE and the von Hippel index. They were 1.004 and 2.86. The value 1.004 is quite close to 1.00 and the percent instability of the standard error was under 3%. The number of imputations I used seems reasonable.

Rounding. Graham (2009) states that rounding of imputed values should be minimized even when the metric of the target variable is binary, discrete, or integer in character. Rounding, it turns out, can bias variance estimates. Research suggests rounding is not helpful unless the analytic method that will ultimately be applied to the data does not accommodate fractional values (e.g., logistic regression where the outcome must be scored 0, 1; one would not want fractional values for the outcome).

For categorical predictors with more than 2 levels, analysts typically use dummy variables to represent the different categories of the variable. With some imputation methods, it is possible to obtain illogical patterns of imputed values (e.g., values greater than 1 for dummy variables that have used 0, 1 coding). Most simulation studies suggest that rounding in such cases still is not a good idea (Allison, 2005; Bernaards et al., 2007; Horton, Lipsitz, & Parzen, 2003; Yucel, He, & Zaslavsky, 2008). For the Bayes, PMM, factored regression, and robust multiple imputation methods discussed in this chapter, rounding issues for binary, nominal and categorical variables are irrelevant, which is yet another reason to prefer these methods to others.

Clustered and Multi-Level Data. For multilevel or clustered data, missing data can occur on the outcome variable, the level 1 predictors, the level 2 predictors, and/or the variable that defines the clusters. When it occurs on the cluster id variable, which is rare, most software, including Mplus, listwise deletes all of the cases in the cluster.

Some researchers impute data in two level models as if they were single level models, ignoring the two level structure of the data. Grund, Lüdtke and Robitzsch (2018) conclude that such imputation strategies should be avoided because of the bias they can introduce.

Mplus Bayesian imputation can be used effectively for two level data. However, the use of H1 imputation with a general covariance structure in multi-level modeling is challenging because of convergence and identification issues. In addition to the conditions described earlier for single level models, the number of variables in your H1 imputation model must be less than the number of clusters. As well, if the number of level 2 variables exceeds the number of clusters, the model will not be identified. Generally speaking, variables with a large number of missing values will tend to cause problems for multiple imputation with multilevel models. You can address identification and convergence problems by reducing the number of variables in the imputation model

but often a better strategy is to use H0 imputation (Asparouhov & Muthén, 2021) or the factored regression approach of BLIMP (Enders, 2022; Enders, Du, Keller, 2020).

Table 26.9 presents Mplus syntax for Bayesian imputation for a multi-level clustered RET conducted in a school setting using an H1 imputation approach (see Chapter 25 for a discussion of clustered RET designs; I assume here you are familiar with the material in that chapter). The cluster variable is classrooms and is in the variable called CLASS. Classrooms are randomly assigned to a treatment versus control condition (the variable TREAT). The treatment addresses three mediators (HWORK1, HWORK2, and HWORK3) to help students approach their math homework more effectively and efficiently. The outcome variable (MATH) is end of the year scores on a math test. Level 2 variables are the grade level (e.g., 7th grade, 8th grade) of students in the class (cgrade) and an index of teacher experience in teaching math (TEACH).

Table 26.9: Two-Level H1 Imputation

```
1. TITLE: TWO LEVEL IMPUTATIONS FOR H1 MODEL ;
2. DATA:  FILE = howework.dat ;
3. VARIABLE:
4. NAMES = class stuid ses cgrade hwork1 hwork2 hwork3 treat
5.  math teach race sctype ;
6. USEVARIABLES = math hwork1 hwork2 hwork3 treat teach cgrade;
7. MISSING ALL(-9999) ;
8. CLUSTER = class;
9. WITHIN  = math hwork1 hwork2 howork3 treat;
10. BETWEEN = teach cgrade;
11. DATA IMPUTATION:
12. IMPUTE = math hwork1 hwork2 howork3 treat teach cgrade ;
13. NDATASETS = 25 ;
14. SAVE = missimp*.dat;
15. ANALYSIS:
16. TYPE = BASIC TWOLEVEL ;
```

The program is the same as that for single level models except for Lines 8 through 10 which specify the cluster variable and level 1 and level 2 variables. Line 16 identifies the data as two level data. No output line is required.

Product Terms. As noted, SEM models sometimes include product terms to test statistical interaction or curvilinearity. I described earlier the complications that product terms create for missing data for FIML due to multivariate normality assumptions. For multiple imputation methods, probably the best approaches for models with product terms are those that use factored regression (BLIMP) or chained equations that allow you to customize the imputation model to include interaction terms. See the document on the

software for MICE and BLIMP on the resources tab of my webpage for illustrations.

When using the chained equation approach with product terms, you will encounter discussions of the use of **passive imputation** versus **just another variable imputation**. Consider the variables X, Z, and their product term XZ. A passive imputation strategy imputes values into X and Z and then multiplies the imputed values to form the product term. The just another variable (JAV) strategy specifies X, Z, and XZ as variables to impute and treats XZ as any other variable, allowing it to take on whatever value is imputed into it even if it is not the exact product of X and Z. Research suggests the JAV strategy often is the better of the two approaches (Seaman, Bartlett & White, 2012; von Hippel, 2009), with the exception of logistic models, where Seaman et al. (2012) found that passive imputation using predictive mean matching for the components of the product term performed somewhat better. Also, the JAV approach generally works well only when the missing data are functionally MCAR and when the component X and Z are functionally multivariately normally distributed. Current research suggests that the factored regression Bayesian imputation method by BLIMP in conjunction with passive communication tends to work best for multiple imputation contexts (Enders, 2022).

Concluding Comments on Multiple Imputation. There are numerous multiple imputation methods to choose from. The research literature is filled with conflicting advice as to the best way to approach multiple imputation partly because the tools we now have available are different from what we had a decade ago. I reviewed what I consider to be the most promising methods, including Bayesian imputation (encompassing the H0 and H1 methods in Mplus and factored regression Bayesian imputation in BLIMP), chained equations with predictive mean matching, and robust multiple imputation. A distinguishing feature between several of these methods is the reliance on assumptions about joint variable distributions (usually multivariate normality) versus adaptive methods that work with separate conditional equations for each model variable. No imputation method is perfect and each has its strengths and weaknesses. If I decide I need to use multiple imputation for straightforward RETs without moderation or with moderation that does not require product terms, I lean towards the use of the default H1 imputation method of Mplus. If my analysis model has product terms in which I expect moderate to strong interactions or non-linearity, I lean towards using BLIMP. If I work with variables that are outlier prone, I use robust multiple imputation. Chained equations with predictive mean matching is a good approach for sensitivity checks on the above, although it is of limited use when influential outliers operate. Once the multiple data sets have been generated by one of these methods, then each can be analyzed and combined using Mplus per the programming logic of Table 26.7.

There are other multiple imputation frameworks that I have not covered. These

include resampling methods that rely on jackknifing or bootstrapping (e.g., Mashreghi, Léger & Haziza, 2014; Zhou, Elliott, & Raghunathan, 2016). and fractional weighted imputation (Yang & Kim, 2016), to name a few. Standard bootstrapping methods are not applicable in multiple imputation analysis because one must adapt them to accommodate the uncertainty of the imputations. Two general approaches for bootstrap inference with multiple imputation are plausible. In the first approach, we generate, say, 100 imputed data sets and bootstrap estimation is applied to each of them. In the second approach, 100 bootstrap samples of the original data set (including missing values) are drawn and in each of sample the data are multiply imputed. Both approaches have utility but research tends to favor the latter, with the former yielding inflated (more conservative) standard errors. The methods are computationally intense and not readily accessible to everyday researchers (see Bartlett, 2021; Bartlett & Hughes, 2020; Schomaker & Heumann, 2018).

ADDITIONAL ISSUES IN HANDLING MISSING DATA

In this section, I consider three topics tied to missing data analysis, (a) the role of sample size and the overall proportion of missing data to the viability of the analysis of data with missing values, (b) the occurrence of missing data in longitudinal designs, and (c) item-level missing data for multi-item scales.

Sample Size and the Amount of Missing Data

A common question researchers ask is what is the maximum overall proportion of missing data that methods such as FIML and multiple imputation can be applied to effectively. The answer to this question is complex and is a function of the interaction between sample size, the overall proportion of missing data (OPMD) and, to a lesser extent, the type of missing data mechanism(s) and the patterning of missing data. Statements in the literature often assert that the methods work well for missing data rates up to 50%. The idea is that at some point the degree of missingness is so large that the assumptions of the missing data model drive conclusions more than the true state of affairs. Vach (1994) argues the underlying dynamics are too complex to reasonably answer the simplistic question of the magnitude of missing data rates that modern methods can accommodate: "*It is often supposed that there exists something like a critical missing rate up to which missing values are not too dangerous. The belief in such a global missing rate is rather stupid.*" (p. 113).

Let me give some examples as food for thought. Suppose your data are MCAR, your sample size is 30,000 and the overall percent of missing data is near 90%. Your listwise deleted sample is 3,000. The unbiasedness, statistical power, and margin of errors

(MOEs) for the parameters of interest likely will be satisfactory when FIML is applied to the data and the substantive conclusions you make will likely be similar to the case where you had the complete data. Afterall, the listwise deleted sample is N = 3,000! As long as the sample size is sufficiently large, inferences drawn from the average loglikelihood of FIML based on missing data that are MAR or MCAR should still be valid due to the fact that the estimates, confidence intervals, and test statistics become arbitrarily close to their true values given the large N (Morimoto, 2022).

In a careful mathematical analysis of underlying theorems, Mortimoto (2020) notes that the correspondence between missing data average likelihoods and complete data average likelihoods tend to get closer to one another as sample size increases and as the rate of missing data decreases. According to Mortimoto, the interactive influence of these two factors are key to evaluating the viability of missing data analytic strategies. For example, Madley-Dowd et al. (2019) conducted simulations and concluded that "the proportion of missing data should not be used to guide decisions on multiple imputation." Their simulations, however, used a sample size greater than 10,000 which Mortimoto showed tends to render the proportion of missing data moot for most purposes.

Complications can arise for missing data analysis when sample sizes are small. With N < 100, use of FIML in complex SEM models becomes tenuous, although doing so is not out of the question (see Chapter 28). The use of listwise deletion with small N can be problematic because we usually struggle with low statistical power in small N scenarios. Listwise deletion can exacerbate the problem by deleting cases and not making use of all the information that is available. As noted, predictive mean matching can be problematic with small N because of limited donor pools and the occurrence of dependencies due to repeat donors. Multiple imputation typically relies on linear regression for the imputation model, but small sample sizes often necessitate regression models with few predictors. This, in turn, imposes limits on the imputation model. Bayesian imputation can encounter convergence issues with small sample sizes. In short, there are limitations of all the methods for dealing with missing data with small N.

McNeish (2017) evaluated various missing data methods in a simulation study focused on regression coefficients in a single equation linear model with three predictors for sample sizes of 20, 50, 100, and 250 and missing data rates of 10%, 20%, 30%, or 50% under both MAR and MNAR scenarios. Most of the methods fared poorly for all the sample sizes in terms of inflated Type I errors when the rate of missing data was 50%, although predictive mean matching was consistently conservative rather than liberal. All of the methods except FIML adequately controlled Type I errors for missing data rates of 20% or less for all sample sizes. For FIML, control for missing data rates of 20% or less was adequate for Ns of 50 or greater. Predictive mean matching fared poorly in terms of

parameter bias when the sample size was 50 or less. In general, the other methods fared reasonably well when N was 50 or greater and missing data rates were 20% or less, especially methods that relied on a joint distribution multiple imputation method (in the spirit of the H1 Bayesian approach of Mplus). For additional discussion of sample size and multiple imputations, see von Hippel (2013, 2016).

As noted, one reason FIML performs badly for Type I errors with small N is because it is based on asymptotic theory which requires larger N. With small N, many multiple imputation approaches invoke a t distribution for deriving p values and confidence intervals to correct for this problem per the statistical strategy described in Barnard and Rubin (1999). Unfortunately, this method cannot be applied to complex SEM models because it does not generalize well to all types of SEM coefficients (von Hippel, 2016). von Hippel (2016) suggests a t distribution adjustment that might work well with FIML in SEM with small sample sizes based on a modification of the reference degrees of freedom for the t distribution. For details, see von Hippel (2016).

In sum, it is difficult to state simple rules that allow you to know with certainty if the missing data method you plan to use is appropriate for the RET you are planning or that you have conducted. Probably the best way to determine how different methods will fare is to design your own localized simulation that explicitly takes into account your study conditions. I show you how to do so in Chapter 28.

Longitudinal Data

In longitudinal studies with multiple follow-ups or many assessments of the same variables during treatment, individuals are measured repeatedly over time. Some participants drop out of such studies at a given time point and do not provide data at subsequent assessments, a phenomenon described as a **monotone missing data pattern**. In other cases, a person can provide missing data at one follow-up but be measured again at one or more of the next follow-ups yielding what is called a **nonmonotone missing data pattern**. Specialized methods have evolved for missing data in longitudinal designs (Daniels & Hogan, 2008; Molenberghs & Kenward, 2007) but simulation studies suggest that the multiple imputation methods described above can be effectively applied in longitudinal as well as cross sectional contexts (Graham, 2009; Huque, Carlin, Simpson, et al., 2019). In Chapter 16, I described multiple modeling strategies for longitudinal data, most of which are SEM based. Because these methods often employ maximum likelihood estimation, either FIML or likelihood based variations of FIML have evolved for dealing with missing data in them. For mixed effect models, the modeling works with data in long format rather than wide format so that use of the available information is maximized (De Silva et al., 2017; Welch et al., 2014). See Chapter 16 for elaboration.

Items from a Multi-Item Scale

When using multi-item scales, respondents may have missing data on one or more items of the scale, but not all of them. If you plan to conduct item level analyses, such as factor analysis or the analysis of composite reliabilities, then you should conduct those analyses using one of the principled missing data methods described above (e.g., FIML). However, researchers often conduct analyses only on the total score of the scale for each individual in the form of a single indicator. A common strategy in this situation is to (a) impute the individual's mean of the items he or she completes for the scale (with appropriate reverse scoring) into the missing items, and then (b) calculate a total score for the individual based on these complete data. Note that this strategy is not the same as the across individual mean imputation strategy I described earlier and which should not be used. That strategy imputes a sample mean of an item that is calculated *across individuals* who responded to that item. In the current strategy, we instead impute a response for an individual on a given item that is the mean response calculated *within a given individual* to other items that serve as interchangeable indicators of the variable.

An alternative strategy to the above is to signify the total score for any individual who has not completed all items as "missing" and then use standard FIML or multiple imputation methods to deal with the missing total score.

A third strategy is to use principled single imputation to impute the missing item-level data and then to calculate a total score from that "complete" data, recognizing there will be a small amount of potential error in it by virtue of using only a single imputation for the items and not making adjustments accordingly (unless missing data is considerable).

The relative strengths of these approaches have not been rigorously explored. Schafer and Graham (2002) suggest all three approaches may work well, depending on features of the study and scale. Graham (2009) states that relying on partial responses to form a total score should be reasonable if (1) a high proportion of items have been answered by the individual (not fewer than 50%), (2) the completed items still capture the full content domain of the construct, and (3) the items have relatively high internal consistency (see Chapter 3). Enders (2010) discusses alternative approaches to the above three but they are more specialized and applicable only in limited contexts.

Contrary to Graham (2009), Newman (2014) argues for using the mean of items responded to as an index of the total score no matter how few the number of items the person completes. His guiding principle is to always use information if it is available. As an example, Newman would argue for basing the total score on the response to a single item of a 20 item inventory if that is all a person completed. Single items on multi-item scales often have high levels of unreliability making them less trustworthy than a

composite based on multiple items (see Chapter 3). I personally would not have much faith in using an indicator that I felt is highly unreliable. I tend to agree with Graham and prefer using a principled missing data approach in such cases.

In the final analysis, researchers should consider the broader psychometric context of one's study to make decisions about how to best deal with item-level missingness for multi-item scales. Uniform adoption of simplistic rules like "base the total score on the item responses no matter how many items the person responds to" are ill-advised.

LISTWISE MISSING DATA METHODS REVISITED

In 2014, Paul Allison, a noted expert on missing data analysis, wrote an article titled "Listwise missing data: It's NOT evil." The point of the article was to underscore that even though listwise deletion can have shortcomings, it has acquired a somewhat undeserved bad reputation. Listwise deletion should indeed be used with caution. However, when FIML or multiple imputations are not feasible or are potentially problematic, listwise deletion should not be summarily dismissed.

One disadvantage of listwise deletion relative to FIML or correctly performed multiple imputation is its loss of statistical power with the reduced sample size that results from using it. However, as noted earlier, if your sample size remains reasonably large after listwise deletion, then the loss in power may be non-consequential. Keep in mind the following principle: If data are MCAR, then a reduced sample of listwise deleted cases will be a random subsample of the original sample. If the estimates are unbiased for the complete data, the estimates also will be unbiased for the listwise deleted sample and the standard errors will be as appropriate but because we work with less information.

Interestingly, there are situations where data that are not MCAR will yield unbiased estimates of the unstandardized regression/path coefficients under listwise deletion. Suppose that the complete-data model is a correctly specified regression analysis that regresses Y onto a set of predictors. Suppose further that data are missing on both Y and a given predictor X. Allison notes that listwise deletion will produce unbiased estimates of the unstandardized coefficients even when the data are only MAR or NMAR for X as long as the probability of missing data for the predictors or for Y does not depend on Y* (for a proof, see footnote 1 in Allison, 2002). This also applies to listwise deletion for logistic and count regression, but under even more general conditions (Allison, 2014), namely the listwise deleted coefficient estimates (but not the intercept) will be unbiased as long as Y* is not related to missing data on both the outcome and the predictors but it also can be related to either one alone.

It turns out that listwise deletion is one of the more robust methods to dealing

with violations of MAR among predictors in regression contexts, again, as long as missingness is independent of Y*. To quote Allison (2002):

> *The methods of maximum likelihood and multiple imputation… are potentially much better than listwise deletion in many situations, but for regression analysis, listwise deletion is even more robust than these sophisticated methods to violations of the MAR assumption. Specifically, whenever the probability of missing data on a particular independent variable depends on the value of that variable (and not the dependent variable), listwise deletion may do better than maximum likelihood or multiple imputation.* (p. 7).

Despite the above, listwise deletion can indeed result in biased coefficient estimates when data are not MCAR and it must be used with caution. For examples, see Enders (2010, 2022) and Schafer and Graham (2002)

In sum, with large sample sizes and low missing data conditions, listwise deletion can be effectively used in many (but not all) scenarios. The method sacrifices the additional statistical power that can be gained by using FIML or multiple imputations, but sometimes the amount of increased power is small. Technically, listwise deletion assumes MCAR but there are MAR or NMAR cases where it will work satisfactorily. And, as I noted earlier, if one is using a sample-then-population approach to inference, one can always fall back on assertions that study conclusions apply to populations of individuals who provide complete data. Depending on context, such an assertion may not be problematic if such individuals are representative of the larger population *on the variables that matter*.

WHICH METHOD IS BEST?

As stated at the outset of this chapter, the best method for dealing with missing data is not to have any. To the extent possible, you should pursue methodological strategies that minimize missing data. Given missing data, Allison (2012c) has argued that FIML generally should be the preferred approach to missing data for multiple reasons. FIML is simple to apply, it generally is more efficient (in a statistical sense of the term), its results are reproducible because it does not rely on a random process, and it does not have conflict between the analytic model and the imputation model. For multiple imputation, it generally is important for your analysis model and imputation model to be congenial; the models do not have to be identical but, as Allison (2012c) notes, they can't have major inconsistencies. Finally, FIML does not have to deal with a range of issues that multiple

imputation does, such as rounding of imputed scores and the number of imputation data sets to use. Example studies that have found FIML to be superior to multiple imputation methods include Larsen (2011), Yuan, Wallentin, and Bentler (2012) and von Hippel (2013). However, see Lee & Shi (2021) for exceptions and also consider the small sample research described earlier.

The Bayesian full information missing data approach shares many of the same advantages as FIML and is best if you want to pursue Bayesian modeling for your research questions and you have sufficient sample size. As noted, the Bayesian approach will generally produce results similar to FIML when non-informative priors are used, although, technically, it is a distinct analytic framework.

If FIML is not available for the RET analyses you want to conduct or if you worry about too strong of violations to multivariate normality, then a multiple imputation approach that uses BLIMP or chained equations with tailored imputation models for each variable might be the best way to go. If outliers are of concern, then a robust multiple imputation method can be pursued. A limitation of multiple imputation analysis in Mplus is that you cannot use bootstrapping. This is not true of FIML. FIML and Bayesian full information approaches also require that you have confidence in your fitted model. Multiple imputation can make use of additional predictive variables in the imputation model in ways that are different from FIML and, in this sense, can produce different results/conclusions (Collins, Schafer & Kam, 2001).

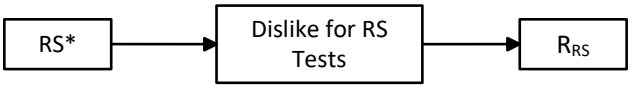
Finally, if none of the above methods are viable and if your sample size is sufficiently large and missing data are few (e.g., less than 15% or so), you might be able to use listwise deletion with the appropriate cautions outlined earlier.

For FIML modeling using SEM, there is controversy about whether to treat exogenous variables as fixed predictors or random predictors, or in the vernacular of Mplus, to formally "bring exogenous variables into the model" by specifying their variances in Mplus syntax. By treating them as fixed, they are listwise deleted and FIML is applied only to the endogenous variables. No distributional assumptions about the exogenous variables are required. If they are "brought into the model" then FIML makes multivariate normality assumptions about them, just as it does any other endogenous variable. FIML often is robust to violations of these assumptions but sometimes it is not. Unfortunately, we have little guidance on what to do regarding the treatment of exogenous variables. I describe below a method for conducting localized missing data simulations that can provide you perspectives for the particular RET you are conducting.

WHEN DATA ARE NOT MCAR NOR MAR

When data are not MCAR or MAR and if assumption violations for MAR are deemed

sufficiently strong, one might consider using modeling methods appropriate for MNAR data. One's first priority, of course, should be to measure the source of missing data dependencies and to statistically adjust for them to remove the offending MNAR dependencies. Suppose in an RET to improve reading skills (RS) I think that students with lower RS will tend to avoid (i.e., miss) the posttest measurement session because they find reading tests frustrating and, hence, dislike them more than students higher on RS. This missing data theory can be represented by the following causal model linking RS* (the reading skill scores of everyone, whether they have missing data or not) to missingness on the observed measure of reading skills, the latter which I symbolize as a binary dummy variable R_{RS} (0 = has missing data on RS, 1 = does not have missing data on RS):



The diagram depicts a scenario in which the data are MNAR because there is a dependency between RS* and R_{RS}. In the diagram, I include a mediator of that dependency in my theory of missingness that I can take advantage of to reduce the undesired dependency. Specifically, if I obtain a measure of liking/disliking of reading tests at baseline, then I can statistically control for it in my modeling of reading skills, either as a direct covariate or perhaps as an auxiliary variable. Suppose when I control for liking/disliking of reading tests, the dependency is reduced but there still remains a residual dependency between RS* and R_{RS}. Such a result suggests that disliking of reading tests is not the sole mediator of the relationship between RS* and R_{RS}, leaving me with a scenario that still is MNAR. Sometimes the amount of such residual bias will be small and can be ignored without consequence; other times it will be non-trivial and must be reckoned with. Collins, Schafer, and Kam (2001) explored different forms of MNAR bias and found that such bias tended to be problematic primarily when the rate of missing data is relatively large (greater than 25%) for sample sizes typical in the social sciences. For rates of 25% or less, failing to control for omitted causes of missing data dependencies produced practically significant bias only when the (residual) dependencies were fairly strong, e.g., dependencies corresponding to correlations of 0.40 or greater.

Methods for analyzing MNAR data in such cases are described in Allison (2002), Graham (2009), Enders (2010, 2022) and Muthén, Muthén and Asparouhov (2016). The two most common strategies are a **pattern mixture modeling approach** and a **selection modeling approach**. The former uses dummy variables corresponding to missing data patterns that occur in the data and it models how these patterns influence the parameters

in the substantive model of interest. In RETs, pattern mixture modeling often results in under-identified models so they are not very practical. As well, both pattern mixture and selection modeling make strong, untestable assumptions whose violation can be problematic in their own right (Demirtas & Schafer, 2003), perhaps more so than just analyzing the data using FIML under incorrect assumptions of MAR (Enders, 2010). Indeed, some methodologists argue against the routine use of pattern mixture and selection modeling other than as sensitivity tests (Allison, 2002; Demirtas & Schafer, 2003). From this vantage point, you compare your results when assuming MAR using FIML with the results you obtain using, say, selection modeling. If the results are similar, you feel more confident in them. If the results differ under the two scenarios, you don't know which set of results to believe and you must live with the uncertainty accordingly.

Given their controversial nature, I do not delve into MNAR analyses here. I include a document on my webpage that describes the basics of selection modeling for MNAR missing data scenarios using Mplus.

MISSING DATA SIMULATIONS

When planning an RET or after an RET has been conducted it can be helpful to conduct a localized Monte Carlo simulation to determine if the missing data analytic strategy you plan to use is workable. I discuss localized simulations in depth in Chapter 28 for sample size determination but they can also be adapted to evaluate the impact of missing data on the quality of statistical inferences. I defer illustration of the approach to Chapter 28 but encourage you to read the material on missing data in that chapter.

CONCLUDING COMMENTS

Missing data are not uncommon in the social and health sciences. One way or another, researchers must decide how to deal with them. The traditional approaches of listwise and pairwise deletion have been replaced by more modern methods, although both of the older methods remain viable under certain conditions, more so listwise deletion than pairwise deletion.

Data can be missing completely at random (MCAR), missing at random (MAR), or it can be not missing at random (NMAR). The extent to which data approximate MCAR and MAR has implications for the choice of a missing data analytic strategy as does the amount of missing data, the assumptions made by the missing data strategy, the pattern of missing data, the broader statistical model that is being used, the sample size, and the type of questions that are being addressed in the context of that model. Given this as well as the plethora of available methods, choices of which strategy to use is often confusing

for applied researchers.

Tests for possible bias (non-randomness) in missing data work with missing data dummy variables or some variant of them. The available tests generally cannot tell us definitively if data are MCAR or MAR, but they can increase our confidence to a greater or lesser extent that these properties hold. Non-randomness is a matter of degree, so data can be "functionally" MCAR or "functionally" MAR even if they are not strictly MCAR or MAR. Also, bias is not necessarily problematic if it is unrelated to the research questions being addressed. Finally, bias can affect more the external validity rather than the internal validity of a study, depending on the sampling model that conceptually guides the research (population-then-sample versus sample-then population).

In general, one will choose among four classes of missing data strategies, (1) listwise deletion, (2) principled single imputation methods, (3) full information estimation methods, and (4) multiple imputation methods. Each has strengths and weaknesses. Recent research tends to somewhat favor FIML methods over multiple imputation methods when both are available, especially for larger sample sizes. Listwise deletion probably works reasonably well as long as the amount of missing data is not too high (say less than about 25%), the data are MCAR, and the sample size is large.

Traditionally, missingness is dealt with in the context of the specific variables included in one's broader statistical model. Some methodologists feel that including additional variables in the analysis that are not of substantive interest but that can inform the dynamics of missingness is of value. The use of such auxiliary variables can be helpful in theory, but simulations that use realistic research scenarios suggest their incremental value may not be as widely beneficial as many believe. I return to this topic in Chapter 27 where I discuss missing data dynamics associated with dropping out of treatments.