1

Introduction to Missing Data

1.1 CHAPTER OVERVIEW

It goes without saying that missing data are a pervasive interdisciplinary problem. Not surprisingly, how we deal with the issue can have a major impact on the validity of statistical inferences and the substantive conclusions from a data analysis. In a highly cited paper nearly 20 years ago, Schafer and Graham (2002) described maximum likelihood estimation and Bayesian multiple imputation as "state-of-the-art" missing data-handling procedures. A lot has changed since then, and these approaches are now considerably more mature and far more capable than they were at the time. The Bayesian paradigm has simultaneously gained in popularity and is now an important alternative to maximum likelihood and multiple imputation rather than an estimation method co-opted for the latter. This trio of contemporary analytic approaches forms the core of the book, which I've rewritten from the ground up to showcase new developments and applications.

Modern missing data-handling procedures have a lot to offer, but we need to understand when and why they work. The first half of this chapter sets the stage with a summary of Rubin and colleagues' theoretical framework for missing data problems (Little & Rubin, 1987, 2020; Mealli & Rubin, 2016; Rubin, 1976). This nearly universal classification system comprises three missing data mechanisms or processes that describe different ways in which the probability of missing values relates to the data. From a practical perspective, Rubin's mechanisms function as data analysis assumptions that dictate the validity of our statistical inferences. As you will see, these assumptions involve mostly untestable propositions, although we can take steps to make certain conditions more plausible. This includes leveraging additional variables that carry information about the missing values but are not part of the main analysis plan.

The middle section of the chapter describes a small selection of older missing data-handling methods. Methodologists have been studying missing data problems for the better part of a century, and the statistical literature is replete with potential solutions, most of which are historical footnotes. Researchers are now broadly aware that better options are available, so I limit this section to a small collection of strategies you may still encounter in published research articles or statistical software packages. I use computer simulation studies to highlight the shortcomings of these methods relative to modern approaches such as maximum likelihood estimation.

The chapter concludes with sections on planned missing data designs that introduce intentional missing values as a device for reducing respondent burden or lowering research costs. Purposefully creating missing data might seem like a bad idea, but this strategy is perfectly appropriate and cannot introduce bias. Although analyzing fewer data points necessarily reduces power, the reduction can be surprisingly small, especially for longitudinal variants of these designs. I describe strategies for creating good designs, and I illustrate how to use computer simulations to vet their power.

1.2 MISSING DATA PATTERNS

A **missing data pattern** refers to the configuration of observed and missing values in a data set. This term should not be confused with a missing data mechanism, which describes possible relationships between the data and one's propensity for missing values. Roughly speaking, patterns describe *where* the holes are in the data, whereas mechanisms describe *why* the values are missing. Figure 1.1 shows six prototypical missing data patterns, with shaded areas representing the location of the missing values. The univariate pattern in panel a has missing values isolated on a single variable. This pattern could occur, for example, in an experimental setting where outcome scores are missing for a subset of participants. A univariate pattern is one of the earliest missing data problems to receive attention in the statistics literature, and a number of classic resources are devoted to this topic (e.g., Little & Rubin, 2020, Ch. 2). Panel b shows a monotone missing data pattern from a longitudinal study where individuals with missing data at a particular measurement occasion are always missing subsequent measurements. Monotone patterns received attention in the early literature, because this configuration of missing values can be treated without complicated iterative estimation algorithms (Jinadasa & Tracy, 1992; Schafer, 1997, pp. 218–238).

The general pattern in panel c has missing values scattered throughout the entire data matrix. Importantly, the three contemporary methods that form the core of this book—maximum likelihood, Bayesian estimation, and multiple imputation—work well with this configuration, so there is generally no reason to choose an analytic method based on the missing data pattern alone. Panel d illustrates a planned missing data pattern where three of the variables are intentionally missing for a large proportion of respondents (Graham, Hofer, & MacKinnon, 1996; Graham, Taylor, Olchowski, & Cumsille, 2006). As described later in the chapter, planned missingness designs can reduce respondent burden and research costs, often with minimal impact on statistical power. Panel e shows a pattern where a latent variable (denoted Y_4^*) is missing for the entire sample. This pattern will surface in Chapter 6 with categorical variable models that view discrete responses as arising from an underlying latent variable distribution (Albert & Chib, 1993; Johnson & Albert, 1999).

One final configuration warrants attention, because it can introduce estimation problems for modern missing data-handling procedures. For lack of a better term, I refer

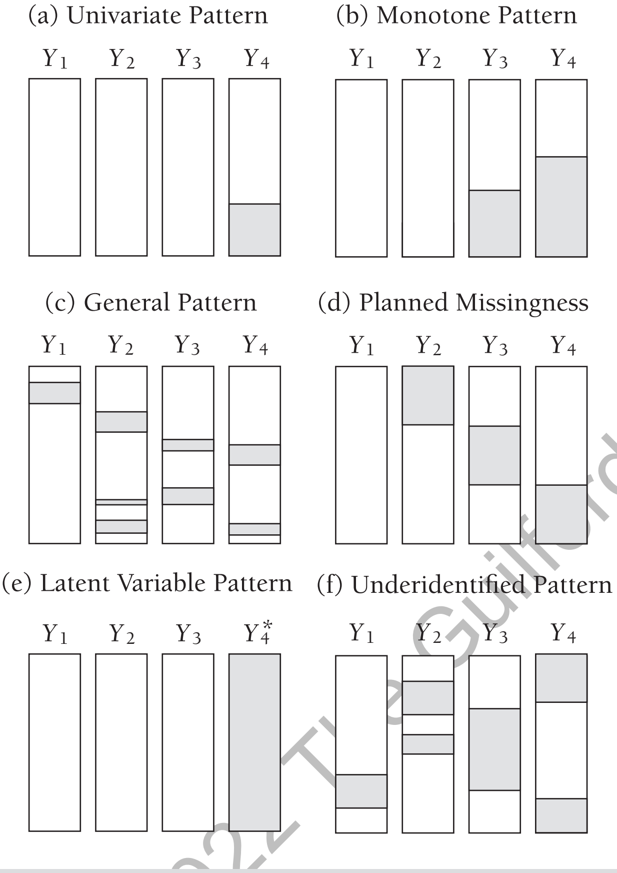

FIGURE 1.1. Six missing data patterns. The gray shaded areas of each bar represent missing observations.

to the configuration in panel f as an *underidentified* missing pattern, because the data provide insufficient support for estimation. The figure depicts a situation where the proportion of cases with data on both Y_3 and Y_4 is so small that it would be difficult or impossible to estimate the bivariate association between these variables. This pattern often occurs with pairs of categorical variables, where unbalanced group sizes and missing data combine to produce very low or even zero cell counts in a cross-tabulation table. It is important to screen for this configuration prior to conducting a missing data analysis.

1.3 MISSING DATA MECHANISMS

Rubin and colleagues (Little & Rubin, 1987; Rubin, 1976) introduced a classification system for missing data problems that is virtually universal in the literature. This work outlines three **missing data mechanisms** or processes that describe different ways in which the probability of missing values relates to the data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). From a practical perspective, these processes are vitally important, because they function as statistical assumptions for a missing data analysis. However, the terms can be confusing (e.g., missing at random refers to a systematic process), and published research articles sometimes conflate their meaning. In the years since Rubin's seminal work, methodologists have clarified certain aspects of his original definitions (Mealli & Rubin, 2016; Raykov, 2011; Seaman, Galati, Jackson, & Carlin, 2013) and have added special subtypes of processes (Diggle & Kenward, 1994; Little, 1995). As an aside, I mostly avoid acronyms throughout the book, but I generally refer to missing data mechanisms by their abbreviations.

Partitioning the Data

Rubin's missing data theory envisions a hypothetically complete data set partitioned into observed and missing components. To illustrate, Table 1.1 shows a data excerpt from a sample of 500 observations and three variables. The complete data in the leftmost set of columns is partly imaginary, because some its values are missing. The would-be scores are shown in bold typeface. The table's middle two sets of columns separate the observed and missing parts of the data. Symbolically, this partition is $\mathbf{Y}_{(\text{com})} = (\mathbf{Y}_{(\text{obs})}, \mathbf{Y}_{(\text{mis})})$, where $\mathbf{Y}_{(\text{com})}$ denotes the hypothetically complete data, $\mathbf{Y}_{(\text{obs})}$ represents the observed scores, and $\mathbf{Y}_{(\text{mis})}$ contains the would-be values of the missing data. Although $\mathbf{Y}_{(\text{com})}$ and $\mathbf{Y}_{(\text{mis})}$ are fairly standard in the literature, other sources use $\mathbf{Y}_{(0)}$ and $\mathbf{Y}_{(1)}$ (Little & Rubin, 2020; Mealli & Rubin, 2016).

The missing data mechanisms described below are essentially models that explain whether a participant has missing values and how those tendencies relate to the realized data in $\mathbf{Y}_{(\text{obs})}$ or $\mathbf{Y}_{(\text{mis})}$. The target of these missingness models is a set of missing data indicators that functions as random variables. We may or may not need to specify

TABLE 1.1. Would-Be Complete Data Partitioned into Observed and Missing Parts

	Complete			Observed			Missing			Indicators		
ID	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3	M_1	M_2	M_3
1	13	30	**15**	13	30	—	—	—	15	0	0	1
2	19	38	28	19	38	28	—	—	—	0	0	0
3	20	18	8	20	18	8	—	—	—	0	0	0
4	**17**	39	**28**	—	39	—	17	—	28	1	0	1
5	22	26	12	22	26	12	—	—	—	0	0	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
496	**14**	36	22	—	36	22	14	—	—	1	0	0
497	28	**12**	7	28	—	7	—	12	—	0	1	0
498	22	30	10	22	30	10	—	—	—	0	0	0
499	24	38	13	24	38	13	—	—	—	0	0	0
500	**29**	**8**	8	—	—	8	29	8	—	1	1	0

distributions for these variables, but they are nevertheless integral to the theory. The rightmost set of columns in Table 1.1 show the matrix of binary missing data indicators **M** that code whether scores are observed or missing; $M_v = 0$ if a participant's score on variable Y_v is observed, and $M_v = 1$ if Y_v is missing.

Missing data mechanisms describe different ways in which the pattern of 0's and 1's may relate to the realized data in $\mathbf{Y}_{(obs)}$ or $\mathbf{Y}_{(mis)}$. Rubin's framework describes three possibilities: The MCAR mechanism stipulates that the propensity for missing values is unrelated to the data; an MAR process posits that missingness is related to the observed parts of the data only; and an MNAR mechanism allows missingness to depend on the unseen scores. To make each mechanism more concrete, I used computer simulation to create bivariate data sets that conform exactly to each process. I modeled the artificial samples after the perceived control over pain and depression variables from the chronic pain data set on the companion website. This data set includes psychological correlates of pain severity (e.g., depression, pain interference with daily life, perceived control over pain) from a sample of $N = 275$ individuals suffering from chronic pain. Figure 1.2 shows the scatterplot of the hypothetical complete data (i.e., $\mathbf{Y}_{(com)}$) for an artificial sample of the same size. The contour rings convey the perspective of a drone hovering over the peak of the bivariate normal population distribution. I subsequently deleted 50% of the depression scores following each mechanism.

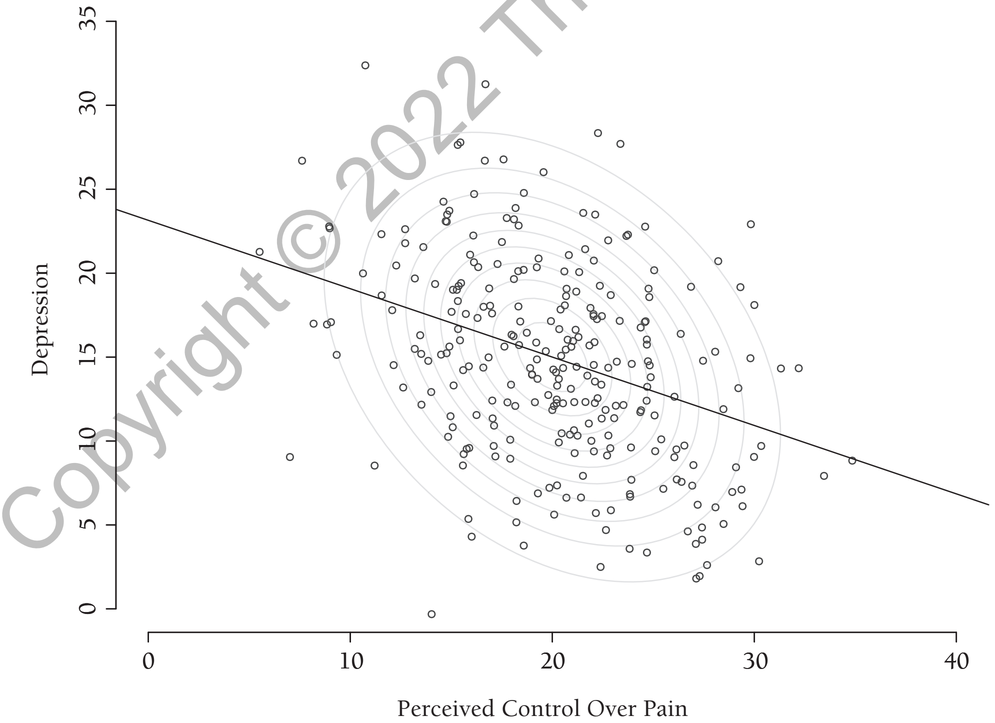

FIGURE 1.2. Complete-data scatterplot showing the would-be values of two variables from a sample of 250 participants.

Missing Completely at Random

A **missing completely at random mechanism** states that the probability of missing values is unrelated to both the observed and missing parts of the realized data. This process is what researchers think of as purely haphazard missingness. The formal definitions of Rubin's mechanisms involve the conditional distribution of the indicator variables in **M** given the realized data in $\mathbf{Y}_{(\text{obs})}$ and $\mathbf{Y}_{(\text{mis})}$. The distribution for an MCAR process is

$$\Pr\left(\mathbf{M}=1\,|\,\mathbf{Y}_{(\text{obs})},\mathbf{Y}_{(\text{mis})},\boldsymbol{\phi}\right)=\Pr\left(\mathbf{M}=1\,|\,\boldsymbol{\phi}\right) \tag{1.1}$$

where $\boldsymbol{\phi}$ is a set of missingness model parameters that link the data to the indicators (e.g., $\boldsymbol{\phi}$ could contain logistic or probit regression coefficients). The left side of the expression, which contains the full complement of possible associations between the indicators and the data, says that the probability of a missing score depends on both the observed and missing parts of the data, as well as some parameters that dictate missingness. The MCAR process on the right side of the expression simplifies by eliminating all dependence on the realized data. In other words, the equation says that all participants have the same chance of missing values, and the parameters in $\boldsymbol{\phi}$ define the overall probabilities of missing data.

A directed acyclic graph is a useful graphical tool for representing the missing data mechanism in Equation 1.1 (Mohan, Pearl, & Tian, 2013; Thoemmes & Mohan, 2015). Figure 1.3a depicts an MCAR process involving a complete variable, X, an incomplete variable, Y, and a binary missing data indicator, M_Y. The white circle labeled Y represents the hypothetically complete variable (i.e., the combination of $Y_{(\text{mis})}$ and $Y_{(\text{obs})}$), and the circle labeled Y^* represents realized values of Y (i.e., $Y^* = Y$ when the missing data

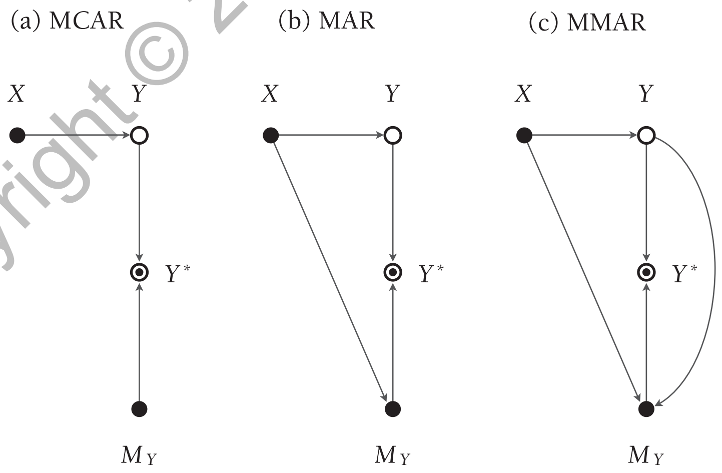

FIGURE 1.3. Directed acyclic graphs that depict missing data processes involving one complete variable, X, one incomplete variable, Y, and a binary missing data indicator, M_Y. The white circle labeled Y represents the hypothetically complete variable, and the circle labeled Y^* denotes the realized values of Y.

indicator $M_Y = 0$ and is missing whenever $M_Y = 1$). Two features of the graph convey an MCAR mechanism. First, the absence of arrows pointing to M_Y indicates that all sources of missingness are contained in the indicator and no other variables predict nonresponse. Second, directed acyclic graph rules tell us that the unseen values of Y are unrelated to M_Y, because the $M_Y \rightarrow Y^* \leftarrow Y$ path connecting the two variables is blocked by a third variable with two incoming arrows (Y^* is a so-called "collider variable").

Rubin's missing data mechanisms can further be viewed as distributional assumptions for the missing values. The definition in Equation 1.1 implies that the missing and observed scores share the same overall (marginal) distributions. To illustrate this point, I randomly removed 50% of the artificial depression scores from the complete data set in Figure 1.2 (i.e., missingness was determined by an electronic coin toss). Figure 1.4 shows the scatterplot of the resulting data, with gray circles representing complete cases and black crosshairs denoting partial data records with perceived control over pain scores but no depression values. Figure 1.2 shows that missing scores are unsystematically dispersed throughout the entire distribution, such that the circles and crosshairs completely overlap, with no differences in their center, spread, or association. The graph highlights that the observed data are a simple random sample of the hypothetically complete data set.

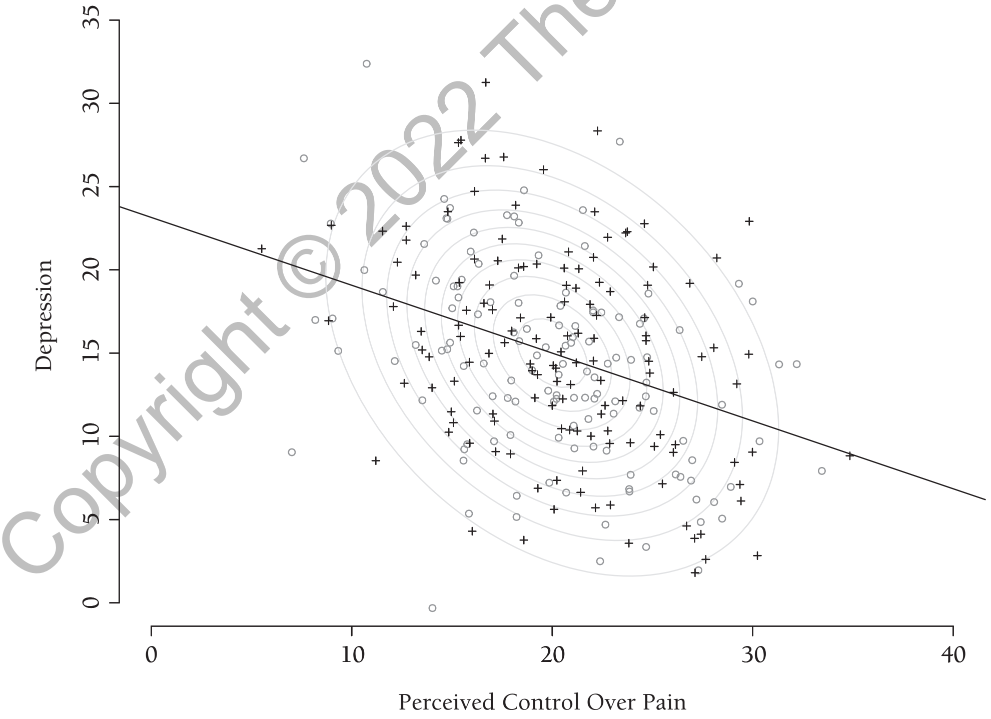

FIGURE 1.4. Scatterplot showing an MCAR process where 50% of the scores are missing haphazardly in a way that does not depend on the data. Circles denote complete observations, and crosshairs denote pairs with missing depression scores.

Missing at Random

A **missing at random mechanism** states that the probability of missing values is related to the observed but not the missing parts of the realized data. The formal definition of this process is as follows:

$$\Pr\left(\mathbf{M}=1 \mid \mathbf{Y}_{(\text{obs})}, \mathbf{Y}_{(\text{mis})}, \boldsymbol{\phi}\right)=\Pr\left(\mathbf{M}=1 \mid \mathbf{Y}_{(\text{obs})}, \boldsymbol{\phi}\right) \tag{1.2}$$

The right side of the equation says that the would-be scores in $\mathbf{Y}_{(\text{mis})}$ carry no additional information about missingness above and beyond that in the observed data. The term *missing at random* is often misunderstood, because it seems to imply a haphazard process instead of a systematic one. Rather, the phrase means that missingness is purely random *after conditioning on or controlling for* the observed data. Said differently, two participants with identical observed score profiles would share the same chance of missing values, whereas two participants with different observed score profiles would have different missingness rates. To clarify this idea, Graham (2009) refers to this mechanism as **conditionally missing at random** (CMAR), and I often do so as well.

The directed acyclic graph in Figure 1.3b depicts an MAR process that features a directed arrow from X to M_Y. The graph shows that the unseen values in Y are *potentially* related to missingness via the $M_Y \leftarrow X \rightarrow Y$ path (in the parlance of this graphical framework, Y and M_Y are said to be *d*-connected). Graphing rules further tell us that conditioning on X eliminates the association between Y and M_Y (i.e., satisfies a conditionally MAR process) by closing the $M_Y \leftarrow X \rightarrow Y$ path. Procedurally, conditioning on X means that the missing data-handling procedure leverages all available data, including the partial records for observations with missing Y values. The three analytic pillars of this book—maximum likelihood, Bayesian estimation, and multiple imputation—do just that.

To further illustrate an MAR mechanism, I deleted 50% of the artificial depression scores in Figure 1.2 following a process where the chance of a missing value increased as perceived control over pain decreased (e.g., participants with little control over their pain were more likely to experience pain-related disruptions that could lead them to drop out of the study). The selection process was relatively strong, with the predicted probability of missing data increasing from about 16% at one standard deviation above the perceived control mean to 84% at one standard deviation below the mean. Figure 1.5 shows the scatterplot of the data, with gray circles again representing complete cases and black crosshairs denoting partial data records with perceived control scores but no depression values. The figure clearly depicts a systematic process where missing scores are primarily located on the left side of the contour plot. Unlike Figure 1.4, the gray circles (cases with complete data on both variables) are no longer representative of the hypothetically complete data, because there are too many scores at the high end of the perceived control distribution and too few at the low end.

An MAR mechanism can also be viewed as a distributional assumption for the missing values. The definition in Equation 1.2 implies that the observed and unseen values of a variable share the same distribution after controlling for the observed values of other variables (i.e., the two sets of scores follow the same *conditional distribution*).

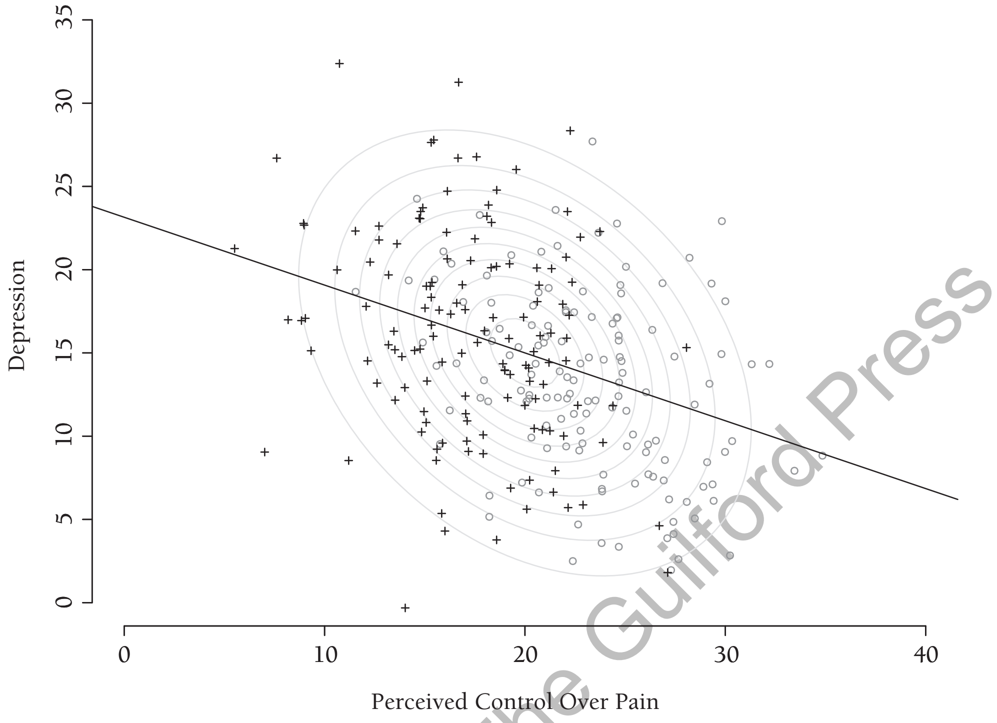

FIGURE 1.5. Scatterplot showing an MAR process where 50% of the depression scores are missing for participants with low perceived control over pain values. Circles denote complete observations, and crosshairs denote pairs with missing depression scores.

Applied to the bivariate normal data in Figure 1.5, this assumption stipulates that the observed and missing depression scores are normally distributed around points on the regression line and share the same constant variation (i.e., the depression distribution is the same for any two individuals with the same perceived control over pain score, regardless of whether they have missing data). Visually, this feature is evident by the fact that the circles and crosshairs lock together like puzzle pieces around the regression line from the hypothetically complete data.

Viewing the MAR process as a distributional assumption provides intuition about the inner workings of contemporary analytic procedures. Although they do so in different ways, maximum likelihood, Bayesian estimation, and multiple imputation all attempt to infer the location of the missing values based on the corresponding observed data. Consider the task of imputing a missing depression score. Given a suitable estimate of the regression line, the MAR process implies that imputations can be sampled from normal distributions centered along the regression line. To illustrate, Figure 1.6 shows the distribution of plausible imputations at three values of perceived control over pain. Candidate imputations fall exactly on the vertical hashmarks, but I added horizontal jitter to emphasize that more scores are located at higher contours near the regression line. Randomly selecting one of the circles from each distribution generates an imputed depression score (technically, imputations are not restricted to the circles displayed in the graph and could be selected from anywhere in the normal distribution, but you get the idea). In fact, Bayesian estimation and multiple imputation both invoke an iterative version of this exact procedure.

Finally, an MAR process is very general and readily extends to multivariate data, although it is more awkward to think about in this context. Returning to the data in Table 1.1, the mechanism must be viewed on a pattern-by-pattern basis. Considering the first row of data (and all other rows where only Y_3 is missing), an MAR process requires that Y_3's missingness is fully explained by Y_1 and Y_2. Moving to the fourth row of data, the mechanism requires that the likelihood of a pattern where Y_1 and Y_3 are both missing depends only on Y_2. Notice that this condition contradicts the statement for the first row, which allows missing values on Y_3 to depend on Y_1. As a final example, the mechanism requires the chance of missing both Y_1 and Y_2 (the pattern in the bottom row of the table) to depend only on Y_3. Again, parts of this proposition are at odds with conditions that govern other patterns. Despite its somewhat clunky construction with multivariate data, Little and Rubin (2020, p. 23) argue that a MAR process is a better approximation to reality than the simpler MCAR mechanism.

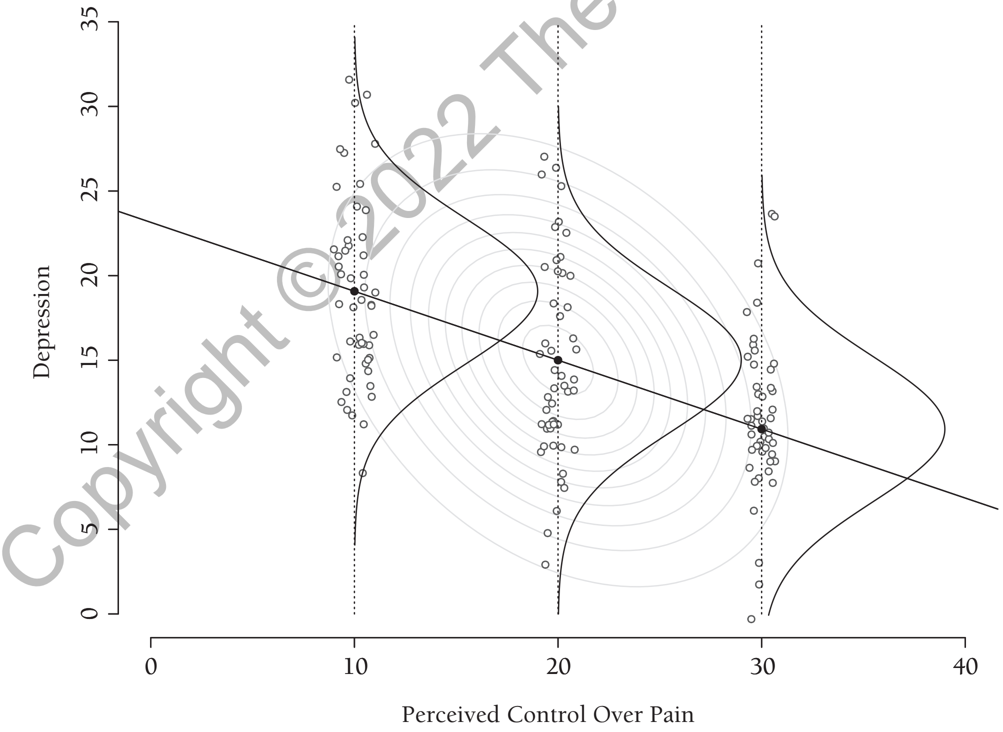

FIGURE 1.6. Distributions of plausible depression imputations at three values of perceived control over pain. Candidate imputations fall exactly on vertical hashmarks, but I added horizontal jitter to emphasize that more scores are located near the regression line.

Missing Not at Random

A **missing not at random mechanism** (also referred to as a not missing at random process) states that the probability of missing values is related to the observed and missing parts of the data. The formal definition of this mechanism is as follows.

$$\Pr\left(\mathbf{M}=1\,|\,\mathbf{Y}_{(\text{obs})},\mathbf{Y}_{(\text{mis})},\boldsymbol{\phi}\right) \tag{1.3}$$

Unlike the previous expressions, the conditional distribution of the missing data indicators doesn't simplify and features two distinct determinants of missingness. Under such a process, two participants with identical observed score profiles no longer have the same chance of a missing value, as the would-be scores themselves carry additional information above and beyond the observed data. Gomer and Yuan (2021) refer to Equation 1.3 as **diffuse MNAR**, because missingness depends on both components of the hypothetically complete data, and they define a **focused MNAR** process as one that depends only on the unseen values in $\mathbf{Y}_{(\text{mis})}$.

$$\Pr\left(\mathbf{M}=1\,|\,\mathbf{Y}_{(\text{mis})},\boldsymbol{\phi}\right) \tag{1.4}$$

Although there is no way to differentiate MNAR subtypes from the observed data, the authors argue that the distinction is important, because diffuse and focused processes can differentially impact one's analysis results. I return to this issue in Chapter 9.

The directed acyclic graph in Figure 1.3c depicts a (diffuse) MNAR process involving the same variables as before. The graph suggests that the unseen values in Y are potentially related to missingness via the $M_Y \leftarrow X \rightarrow Y$ path and the $Y \rightarrow M_Y$ path. As before, conditioning on X closes the $M_Y \leftarrow X \rightarrow Y$ path, thereby eliminating part of the association between Y and its missingness indicator. However, the would-be values of Y still influence missingness via their direct pathway to M_Y. Graphing rules tell us that a pair of connected variables adjacent in a chain cannot be separated, so conditioning on the observed data does not eliminate the dependence between Y and its missing data indicator.

To further illustrate an MNAR mechanism, I deleted 50% of the artificial depression scores in Figure 1.2 following a process where participants with higher levels of depression were more likely to have missing values (e.g., those with acute symptoms would leave the study to seek treatment elsewhere). The selection process was relatively strong, with the predicted probability of missing data increasing from about 16% at one standard deviation below the depression mean to 84% at one standard deviation above the mean. Figure 1.7 shows the scatterplot of the data, with gray circles again representing complete cases and black crosshairs denoting partial data records with perceived control scores but no depression values. The figure illustrates a systematic process where missing scores are primarily located in the top half of the contour plot above the regression line. The gray circles (cases with complete data on both variables) are clearly unrepresentative of the hypothetically complete data.

Unlike the conditionally MAR mechanism, which stipulates that the observed and missing scores share the same distribution after controlling for other variables,

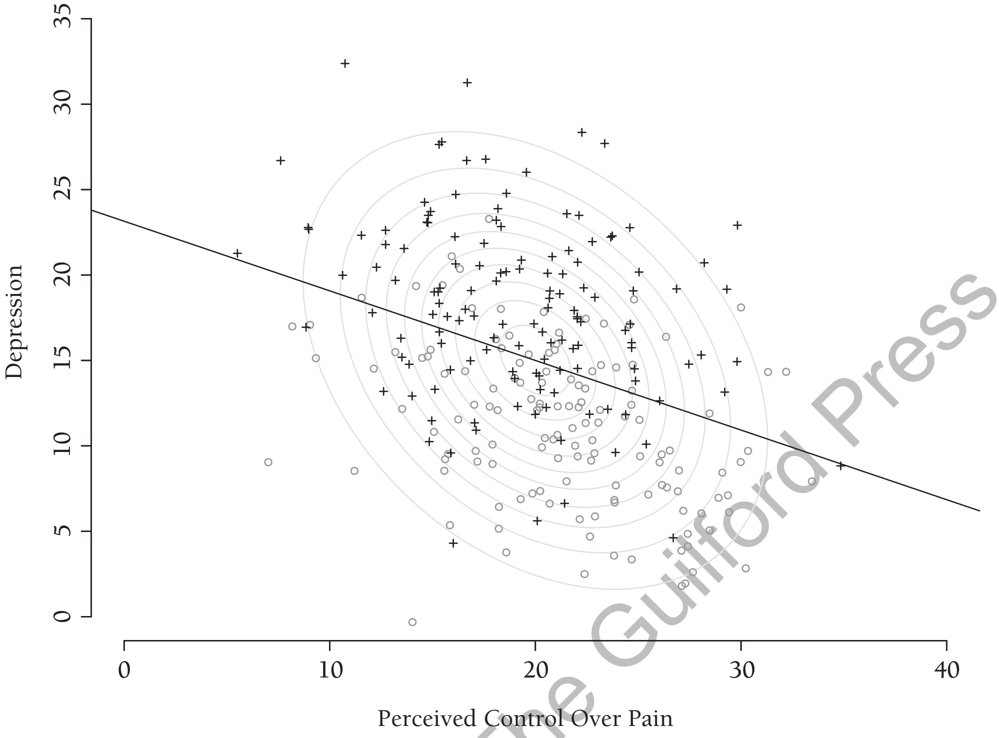

FIGURE 1.7. Scatterplot showing an MNAR process where 50% of the depression scores are missing for participants with high depression. Circles denote complete observations, and crosshairs denote pairs with missing depression scores.

an MNAR process implies that the two sets of scores have different distributions. This situation is clear in Figure 1.7, where the vast majority of the missing scores are above the regression line, and the complete data are mostly below the line. This feature makes imputation considerably more difficult, because there are no data with which to estimate the unique parameters of the missing data distribution. For example, leveraging the perceived control over pain scores alone would create imputations that fall on *either* side of the regression line, and there is no way to formulate an appropriate adjustment without knowing the unseen depression values. As you will see, analytic procedures for MNAR processes (e.g., selection models or pattern mixture models) can only counteract this indeterminacy by invoking relatively strong assumptions about the unseen data.

Mechanisms and Inference

A subtle nuance about Rubin's mechanisms is that they describe missingness *in a specific data set;* that is, the indicators in **M** are fixed at their realized values, and the definitions make no reference to missingness patterns or observed data that could arise from different samples. Rubin's (1976) seminal work clarifies that an MAR mechanism is necessary for obtaining valid maximum likelihood estimates (the same is true for Bayesian estimation and multiple imputation), but this conclusion does not hold for standard errors and significance tests that rely on the frequentist framework and repeated sampling arguments (Kenward & Molenberghs, 1998; Savalei, 2010).

Getting accurate measures of uncertainty under a particular process requires a stricter assumption that the same missingness process *always* generates data sets. Returning to Equation 1.2, valid inferences require the MAR definition to hold for *any* $\mathbf{Y}_{(\text{obs})}$ that you could have worked with, not just the $\mathbf{Y}_{(\text{obs})}$ in a particular sample of data. Statisticians refer to this condition as **missing always at random** (Bojinov, Pillai, & Rubin, 2020; Mealli & Rubin, 2016) or **everywhere missing at random** (Seaman et al., 2013), and Mealli and Rubin (2016) define parallel conditions for MCAR and MNAR processes known as *missing always completely at random* and *missing not always at random,* respectively. Because missingness mechanisms are so prevalent throughout the book, I refer to them by their simpler monikers, with the understanding that measures of uncertainty and significance tests require slightly different definitions.

Ignorable and Nonignorable Missingness

The terms **ignorable** and **nonignorable** missingness are often used synonymously to refer to conditionally MAR and MNAR processes, respectively. In fact, these terms have a somewhat broader definition, although the distinction is relatively unimportant in practice. Rubin's classification scheme features two models: the focal analysis model you would have estimated had the data been complete, and a model that describes the missingness mechanism. These models have parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively. The parameters in $\boldsymbol{\phi}$, whatever they happen to be, are essentially a nuisance, because they are unrelated to the substantive research goals. A key question is, in what situations can we simply estimate $\boldsymbol{\theta}$ from the observed data without worrying about or estimating the missingness model and the parameters in $\boldsymbol{\phi}$? This is the essence of ignorability.

The missingness model is said to be **ignorable** if (1) the missing values follow a conditionally MAR process, and (2) the nuisance parameters in $\boldsymbol{\phi}$ carry no information about the focal parameters in $\boldsymbol{\theta}$ (i.e., $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are **distinct**). Bayesian analyses further require that the two models have independent prior distributions. As mentioned previously, the missing data indicators in **M** function as random variables that follow a distribution. The left side of the equation below is a shorthand way of writing the joint (multivariate) distribution of the observed data and the missing data indicators.

$$f\left(\mathbf{Y}_{(\text{obs})},\mathbf{M}\,|\,\boldsymbol{\theta},\boldsymbol{\phi}\right)=f\left(\mathbf{M}\,|\,\mathbf{Y}_{(\text{obs})},\boldsymbol{\phi}\right)\times f\left(\mathbf{Y}_{(\text{obs})}\,|\,\boldsymbol{\theta}\right) \tag{1.5}$$

I use generic function notation $f(\cdot)$ throughout the book to represent distributions in the abstract without specifying their type or form (e.g., "*f* of something" could be a normal curve, a binomial distribution). If the parameters in $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are independent, applying rules of probability gives the factorization on the right side of the equation. The missingness model is ignorable in this case, because $f(\mathbf{M}|\mathbf{Y}_{(\text{obs})}, \boldsymbol{\phi})$ functions as a constant, and estimating the focal model parameters from the observed data gives the same results with or without this term. In contrast, the missingness model is said to be **nonignorable** if the missing values follow an MNAR process or the nuisance parameters in $\boldsymbol{\phi}$ carry information about the focal parameters in $\boldsymbol{\theta}$. In this situation, we can only get valid estimates of $\boldsymbol{\theta}$ by pairing the focal analysis model with an additional model for missingness (see Chapter 9).

Ignorability is ultimately something we just take on faith, because there is no way to evaluate either of its propositions. Referring to distinctness, (Schafer, 1997, p. 11) says, "In many situations this is intuitively reasonable, as knowing θ [the focal model's parameters] will provide little information about ξ [the missingness model's parameters] and vice-versa." The MAR part of the assumption can be a bit trickier. Van Buuren (2012, p. 33) warns that "the label 'ignorable' does not mean that we can be entirely careless about the missing data," and he goes on to emphasize that satisfying this assumption requires the missing data-handling procedure to condition on all the important determinants of missingness. The next three sections address this point in more detail.

1.4 DIAGNOSING MISSING DATA MECHANISMS

Unfortunately, the observed data do not contain the necessary information to evaluate a conditionally MAR or MNAR mechanism, because both make propositions about the *unseen* scores—the former says the would-be values are unrelated to missingness after conditioning on the observed data, and the latter says they are related. Although methodologists have proposed various diagnostic procedures for evaluating these conditions (Bojinov et al., 2020; Potthoff, Tudor, Pieper, & Hasselblad, 2006; Yuan, 2009a), the validity of contemporary missing data-handling procedures ultimately relies on untestable assumptions and our own expert knowledge about the data and possible reasons for missingness. This leaves an unsystematic MCAR process as the only mechanism with testable propositions.

In truth, evaluating whether missingness is consistent with an unsystematic process isn't necessarily useful, because contemporary methods do not require this strict assumption, and finding that haphazard missingness is (or is not) plausible does not change the recommendation to use these approaches. To this point, Raykov (2011, p. 428) suggests that "the desirability of the MCAR condition has been frequently overrated in empirical social and behavioral research," and I couldn't agree more. Nevertheless, the logic of evaluating an MCAR process warrants brief discussion, because applications of MCAR tests abound in published research articles, and it is important to understand what these tests do and do not tell us about the missing data.

As explained previously, an MCAR process implies that missing and observed scores share the same overall (marginal) distributions; that is, even without conditioning on the observed data, the observed and would-be scores have identical means, variances, and associations with other variables. Kim and Bentler (2002) refer to this condition as *homogeneity of means and covariances*. Methodologists have proposed numerous procedures for evaluating the MCAR mechanism (Chen & Little, 1999; Jamshidian & Jalal, 2010; Kim & Bentler, 2002; Little, 1988b; Muthén, Kaplan, & Hollis, 1987; Park & Lee, 1997; Raykov & Marcoulides, 2014), most of which involve comparing features of the observed data across different missing data patterns. I focus on two simple approaches that consider group mean differences, as these methods enjoy widespread use and are readily available in statistical software.

Univariate Pattern Mean Differences

Perhaps the simplest way to check for an unsystematic process is to form groups of cases with observed or missing scores on a variable Y_v and examine mean differences on other variables (Dixon, 1988; Raykov & Marcoulides, 2014). For lack of a better term, I refer to this as the *pattern mean difference* approach. Returning to the hypothetical data in Table 1.1, this method compares whether the M_1 groups differ on Y_2 or Y_3, the M_2 groups differ with respect to Y_1 or Y_3, and the M_3 groups differ on Y_1 or Y_2.

Returning to the artificial data in Figure 1.4, the pattern mean difference approach creates a missing data indicator that codes whether depression scores are missing or observed and compares the perceived control over pain group means. The $n_{(\text{obs})} = 134$ observations with depression scores had a mean of $M_{(\text{obs})} = 20.08$, and the $n_{(\text{mis})} = 141$ cases with missing data had an average of $M_{(\text{mis})} = 20.52$. This difference equates to less than one-tenth of a standard deviation unit, which is well below Cohen's (1988) small effect size benchmark of $|d| > 0.20$. Because I created missing values by randomly deleting half the scores, it isn't surprising that the mean difference is nonsignificant, $t(273) = .71$, $p = .48$. Raykov (2011) explains that the absence of group differences is necessary but insufficient for demonstrating a purely random process. As such, a safe interpretation is that the data do not contain evidence that refutes the MCAR mechanism.

It is instructive to apply the pattern mean difference approach to data that are not MCAR. Returning to the artificial data in Figure 1.5, participants with low perceived control over pain were more likely to have missing depression scores. In this case, observations with and without depression scores have a perceived control mean of $M_{(\text{obs})} = 23.27$ and $M_{(\text{mis})} = 17.19$, respectively. This difference is equivalent to 1.43 standard deviation units (well above Cohen's large effect size benchmark of $|d| > 0.80$) and is statistically significant, $t(273) = -11.82$, $p < .001$. Importantly, the significant difference implies there is evidence against a purely unsystematic process, but it says nothing about whether a conditionally MAR process is plausible. As a final example, reconsider the MNAR mechanism from Figure 1.7, where participants with elevated depressive symptoms were more likely to have missing depression scores. Despite a very different underlying process, the pattern mean difference is significant and in the same direction, $M_{(\text{obs})} = 21.47$ versus $M_{(\text{mis})} = 19.13$, $t(273) = -3.80$, $p < .001$. This example highlights that the observed data cannot differentiate MAR and MNAR processes. A significant group mean difference implies there is evidence against an MCAR process and nothing more.

Significance tests of pattern mean differences come with a few important caveats. First, a large data set with many variables can yield a staggering number of tests, and correlations among variables allow a univariate difference to masquerade as several significant comparisons. Raykov, Lichtenberg, and Paulson (2012) outline a multiple comparison procedure for this situation, and the multivariate tests of group differences are another option for mitigating false flags (Kim & Bentler, 2002; Little, 1988b). Second, significance tests often suffer from very low power, making them dubious tools for arguing in favor of an unsystematic missingness process. In particular, the power of such tests will be at a maximum when a variable has 50% missing data, because its missingness indicator has equal group sizes. Conversely, lower (or higher) missing data rates cause unbalanced group sizes and lower power. To illustrate, consider the conditionally MAR process depicted in Figure 1.5. Achieving power equal to .80 with a 50% missing data rate requires a standardized pattern mean difference of $|d| > 0.34$ or larger (a small effect size). Had I instead deleted 10% of the data (i.e., group sizes of $n_{(\text{obs})} = 247$ and $n_{(\text{mis})} = 28$), the effect size requirement to achieve the same power increases to $|d| > 0.56$ or larger (a medium effect size). Finally, a pattern mean difference does not automatically imply that the variable in question is a source of nonresponse bias, as the variable's correlation with the focal analysis variables also plays an important role (Collins, Schafer, & Kam, 2001). I return to this point in Section 1.5.

Little's MCAR Test

Little (1988b) proposed a multivariate extension of the pattern mean difference approach that simultaneously evaluates mean differences across a set of variables. The test defines *G* groups of cases that share the same missing data pattern, and it computes the arithmetic means of each pattern's observed data. These pattern-specific means are then compared to maximum likelihood estimates of the grand means. Chapter 3 gives a detailed description of maximum likelihood missing data handling, but for now it is sufficient to know that the estimator leverages the entire sample's observed data without discarding any information. Finally, a test statistic uses the maximum likelihood estimate of the variance–covariance matrix to standardize differences between the pattern-specific means and the grand means. The sum of these standardized differences should be relatively small and close to 0 if scores are MCAR.

Little's test statistic is as follows:

$$T_{\text{L}} = \sum_{g=1}^{G} n_g \left(\bar{\mathbf{Y}}_g - \hat{\boldsymbol{\mu}}_g\right)' \hat{\boldsymbol{\Sigma}}_g^{-1} \left(\bar{\mathbf{Y}}_g - \hat{\boldsymbol{\mu}}_g\right) \tag{1.6}$$

where G is the number of missing data patterns, n_g is the number of cases in missing data pattern g, $\bar{\mathbf{Y}}_g$ contains the arithmetic means for that group, and $\hat{\boldsymbol{\mu}}_g$ and $\hat{\boldsymbol{\Sigma}}_g$ contain the rows and columns of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ (the maximum likelihood estimates) that correspond to the observed variables in $\bar{\mathbf{Y}}_g$. The parentheses contain deviations between pattern g's arithmetic averages and the corresponding grand means, and these are squared (and summed) via matrix multiplication. Multiplying by the inverse of the covariance matrix (the matrix analogue of division) standardizes the discrepancies, such that the numerical value of T_{L} is a weighted sum of G squared z-scores. If values are missing completely at random, T_{L} is approximately distributed as a chi-square statistic with $\sum v_g - V$ degrees of freedom, where v_g is the number of observed scores in pattern g, and V is the total number of variables. Consistent with the mean difference approach, a significant test statistic suggests that missingness is not purely random.

To illustrate Little's test, reconsider the conditionally MAR process depicted in Figure 1.5. In practice, the primary motivation for using Little's test is to evaluate a larger number of variables in $\bar{\mathbf{Y}}_g$, but a bivariate application is useful for illustrating the mechanics of the equation. To begin, the maximum likelihood estimates of the grand means and variance–covariance matrix are as follows:

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 20.31 \\ 14.29 \end{pmatrix} \quad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 27.27 & -13.80 \\ -13.80 & 36.15 \end{pmatrix} \tag{1.7}$$

These means are the benchmark against which to compare pattern-specific means. There are just two missing data patterns in this example: $n_{(\text{obs})} = 141$ observations have scores on both variables (i.e., $v_1 = 2$), and $n_{(\text{mis})} = 134$ cases have missing depression scores (i.e., $v_2 = 1$). The pattern-specific arithmetic means for the two groups are as follows:

$$\bar{\mathbf{Y}}_1 = \begin{pmatrix} 23.27 \\ 12.79 \end{pmatrix} \quad \bar{\mathbf{Y}}_2 = \begin{pmatrix} 17.19 \\ \text{NA} \end{pmatrix} \tag{1.8}$$

Substituting the estimates into Equation 1.6 gives the following test statistic:

$$\begin{aligned} T_{\text{L}} &= 141 \times \left(\begin{pmatrix} 23.27 \\ 12.79 \end{pmatrix} - \begin{pmatrix} 20.31 \\ 14.29 \end{pmatrix} \right)' \begin{pmatrix} 27.27 & -13.80 \\ -13.80 & 36.15 \end{pmatrix} \left(\begin{pmatrix} 23.27 \\ 12.79 \end{pmatrix} - \begin{pmatrix} 20.31 \\ 14.29 \end{pmatrix} \right) \\ &+ 134 \times \frac{(17.19 - 20.31)^2}{27.27} = 98.27 \end{aligned} \tag{1.9}$$

If an unsystematic process generated the data, this test statistic should approximate a chi-square statistic with $\sum v_g - V = (2 + 1) - 2 = 1$ degrees of freedom. The test is statistically significant, $T_{\text{L}}(1) = 98.27$, $p < .001$, indicating that the MCAR mechanism is not plausible for these data. In a multivariate application with more than two variables, a significant test statistic indicates that two or more patterns differ, but the test gives no indication about *which* variables might be responsible.

1.5 AUXILIARY VARIABLES

A conditionally MAR mechanism will be our default assumption until Chapter 9. To refresh, this process stipulates that the would-be scores in $\mathbf{Y}_{(\text{mis})}$ are unrelated to whether a participant has missing values *after conditioning on the observed data*. There are at least two ways this assumption could be violated. First, the unseen scores themselves might predict missingness above and beyond the observed data, as in Figure 1.3c. The only way to counteract nonresponse bias in this scenario is to fit a specialized model that pairs the focal analysis with a nuisance model for missingness (e.g., a selection model or pattern mixture model). Alternatively, the unseen scores may be associated with missingness, because the missing data-handling procedure simply failed to condition on certain variables. In this situation, the MAR assumption *could* be satisfied by controlling for additional or different variables. This scenario is not hard to imagine in practice, as real-world data sets often have hundreds of variables, and controlling for every observation is infeasible. For lack of a better term, I refer to this situation as *MNAR by omission*.

To illustrate an MNAR-by-omission process, suppose that the focal analysis model is the linear regression of Y on X:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1.10}$$

Moreover, suppose that the outcome is missing due to another measured variable *A* that also correlates with *Y*. Figure 1.8a shows a directed acyclic graph that depicts theoretical associations among the three variables and the missing data indicator, M_Y. As before, *Y* represents the hypothetically complete variable, and Y^* represents realized values of *Y* (i.e., Y^* equals *Y* when the missing data indicator $M_Y = 0$ and is missing whenever $M_Y = 1$). Graphing rules imply that *Y* is potentially related to missingness via two pathways: $M_Y \leftarrow X \rightarrow Y$ and $M_Y \leftarrow A \rightarrow Y$.

As explained previously, directed acyclic graphs clarify that conditioning on or controlling for the middle variable in a path eliminates the dependency between the two outer variables. The regression model conditions on *X* and therefore eliminates part of the association between *Y* and M_Y by closing the $M_Y \leftarrow X \rightarrow Y$ path. However, *Y* and M_Y are still related via the $M_Y \leftarrow A \rightarrow Y$ path, so the analysis induces an MNAR-by-omission process, because it fails to condition on *A*. Whether the open path introduces substantial bias depends the magnitude of the association between *A* and M_Y and *A* and *Y* (Collins et al., 2001), but the analysis is nevertheless at odds with the MAR assumption.

Perhaps the simplest way to condition on *A* is to simply include it as an additional covariate in the analysis model as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_i + \varepsilon_i \tag{1.11}$$

This analysis is consistent with an MAR process, because it eliminates all sources of dependency between *Y* and M_Y. However, the model achieves this desirable status by

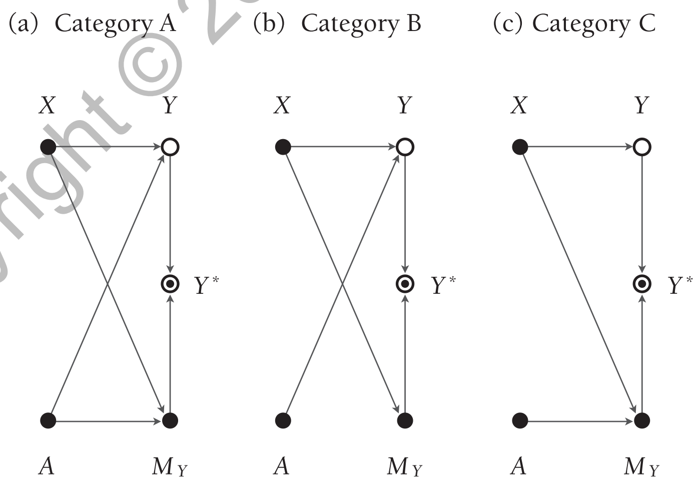

FIGURE 1.8. Directed acyclic graphs that depict missing data processes involving one complete variable, *X*, one incomplete variable, *Y*, a binary missing data indicator, M_Y, and an auxiliary variable *A*. The white circle labeled *Y* represents the hypothetically complete variable, and the circle labeled Y^* denotes the realized values of *Y*.

modifying the meaning of a focal parameter—the β_1 coefficient is a now partial slope that reflects the net influence of *X* above and beyond that of *A*, a variable that wasn't slated to appear in the analysis had the data been complete. Chapters 3 and 5 describe better ways to condition on *A* that don't involve modifying the focal analysis model, but this example nevertheless highlights the importance of conditioning on variables that may not be part of the original analysis plan.

Inclusive Analysis Strategy

The possibility of an MNAR-by-omission process has prompted methodologists to recommend a so-called **inclusive analysis strategy** that introduces auxiliary variables into the focal analysis model or into the imputation process (Collins et al., 2001; Rubin, 1996; Schafer, 1997; Schafer & Graham, 2002). An **auxiliary variable** is an extraneous variable that carries important information for missing data handling but is not part of the focal analysis (or analyses). Conditioning on such variables can fine-tune a missing data analysis, either by reducing nonresponse bias or improving precision. Collins et al. (2001) categorize candidate variables into three buckets: variables that (1) correlate with an analysis variable *Y* and its missingness indicator M_Y, (2) correlate with an analysis variable but not its missingness indicator, and (3) correlate with the missing data indicator but not the analysis variable. The directed acyclic graphs in Figure 1.8 depict these patterns of associations.

The number of variables in many data sets is often so large that an overinclusive strategy is not viable. Reducing a large set of candidate auxiliary variables into one or two principal components is one way to attack this problem (Howard, Rhemtulla, & Little, 2015), but a more tailored approach that selects a small handful of variables often works just as well. Conditioning on category A variables like the one in Figure 1.8a is the top priority, because doing so can improve power and reduce nonresponse bias that results from an MNAR-by-omission process. Moreover, preference should be given to auxiliary variables with the strongest *semipartial* correlations, as variables that account for *unique* variation in the missing variables have the most to offer. Next, conditioning on category B auxiliary variables does not affect bias, but it can improve power by leveraging additional sources of correlation. Again, an auxiliary variable's semipartial association with the incomplete variables is more important than its bivariate correlation. Finally, conditioning on category C auxiliary variables offers no benefits at all. It might seem counterintuitive that ignoring a correlate of missingness (e.g., a variable that exhibits a pattern mean difference) doesn't introduce bias, but the directed acyclic graph in Figure 1.8c clarifies that an MNAR-by-omission process isn't possible, because *A* is not located on a path that connects *Y* to M_Y. The figure reinforces my earlier statement that a pattern mean difference doesn't necessarily signal a source of nonresponse bias.

The utility of an auxiliary variable ultimately boils down to the magnitude of its semipartial correlations with the incomplete analysis variables, as failing to condition on extra variables with weak correlations is unlikely to introduce bias, nor is including such variables going to replace a meaningful amount of missing information. Raykov and West (2015) described a latent variable modeling approach to estimating semipartial correlations with a set of candidate auxiliary variables. Of course, any general-purpose statistical software application can estimate these associations, but most do so after discarding incomplete data records. The advantage of Raykov and West's approach is that it leverages maximum likelihood missing data handling (or alternatively, Bayesian estimation). Again, we don't know how maximum likelihood estimation works yet, but for now it is sufficient to know that the estimator leverages the entire sample's observed data without discarding any information.

A respectable semipartial correlation signals an auxiliary variable that contains unique information about the missing values above and beyond that already contained in the analysis. How large does this correlation need to be in order to reap the benefits of conditioning on the additional variable or suffering the consequences of ignoring it? Simulation studies in Collins et al. (2001) provide some insights. The Collins et al. article examined auxiliary variables with semipartial correlations equal to .32 and .72. Not surprisingly, failing to condition on a variable with a very strong correlation usually produced a bias-inducing MNAR-by-omission process. In contrast, ignoring a variable with the smaller correlation often gave acceptable parameter estimates with little to no bias. Based on these results, it seems reasonable to focus on auxiliary variables with semipartial correlations at least as strong as Cohen's (1988) medium effect size benchmark of ±0.30. Fortunately, we don't need to be too discerning about this cutoff, because these simulations showed no serious consequences of overfitting with a large set of uncorrelated variables. Nevertheless, limiting the number of auxiliary variables is often necessary in practice, because modeling strategies for introducing these extra variables can be prone to convergence failures (e.g., the saturated correlates model; Graham, 2003).

Finally, although the literature has long favored an inclusive strategy (Collins et al., 2001; Rubin, 1996; Schafer, 1997; Schafer & Graham, 2002), it is hypothetically possible that conditioning on an auxiliary variable could enhance rather than reduce nonresponse bias. This could happen, for example, if an auxiliary variable's correlation with an analysis variable and its missingness indicator is fully explained by an unmeasured latent variable. It is unclear how often the constellation of associations needed to cause this problem actually occurs in practice, but interested readers can find an illustration of this phenomenon in Thoemmes and Rose (2014).

1.6 ANALYSIS EXAMPLE: PREPARING FOR MISSING DATA HANDLING

In practice, assuming a conditionally MAR process is usually a good starting point, because this mechanism is more realistic than a purely unsystematic one. Moreover, the three pillars of this book—maximum likelihood, Bayesian estimation, and multiple imputation—naturally leverage this assumption by default. This section serves as a bookend that integrates previous ideas and illustrates two steps to prepare for an MAR-based missing data analysis: comparing participants with and without missing data, and identifying potential auxiliary variables.

To provide a substantive context, I use the chronic pain data on the companion website to illustrate a regression analysis with missing data. The data set includes psychological correlates of pain severity (e.g., depression, pain interference with daily life, perceived control) for a sample of N = 275 individuals with chronic pain. Because the missing data mechanism is an assumption for a specific analysis, I build the example around a linear regression model where depressions scores are a function of pain interference with daily life activities and a binary severe pain indicator (0 = *no, little, or moderate pain,* 1 = *severe pain*).

$$DEPRESS = \beta_0 + \beta_1 \left(INTERFERE_i \right) + \beta_2 \left(PAIN_i \right) + \varepsilon_i \tag{1.12}$$

Approximately 7.3% of the binary pain ratings are missing, and the missing data rates for the depression and pain interference scales are 13.5 and 10.5%, respectively. I use these same variables in Chapter 10 to illustrate missing data handling for a mediation analysis, and I incorporate auxiliary variables from this illustration.

Identifying Correlates of Missingness

Researchers routinely use the pattern mean difference approach to explore whether cases with missing values differ from those with observed data. To illustrate the procedure, I created three missing data indicators that code whether the analysis variables are missing. As before, each dummy code equals 0 if a score is observed and 1 if it is missing. There is no need to examine pattern mean differences for variables already in the analysis, because contemporary missing data-handling approaches automatically condition on this information. Instead, I focus on six continuous variables outside the analysis model: age, exercise frequency, anxiety, stress, perceived control over pain, and psychosocial disability (a construct capturing pain's impact on emotional behaviors such as psychological autonomy and communication, emotional stability, etc.). Three of the candidate variables also have missing data, but incomplete auxiliary variables can still be beneficial as long as their scores are mostly observed whenever the analysis variables are missing (Enders, 2008).

Statistical significance tests are not that valuable for this application, because they lack power due to the highly unbalanced group sizes (e.g., based on of $n_{(\text{obs})}$ = 238 and $n_{(\text{mis})}$ = 37, the depression scale requires a standardized mean difference effect size of nearly 0.50 to achieve .80 power). Instead, Table 1.2 gives the standardized mean difference effect size for each indicator and auxiliary variable. The pain severity indicator produced three comparisons that exceeded Cohen's (1988) small effect size benchmark of ±0.20 (exercise frequency, anxiety, and stress), and the depression indicator produced a single difference of this magnitude (anxiety). Researchers often use logistic regression to predict missingness indicators from study variables, so I also applied this procedure to the example. The logistic analyses further revealed that the set of auxiliary variables explained about 3–4% of the variation in the severe pain and depression indicators, with the anxiety scale producing the largest partial slope.

Before going further, it is useful to step back and take stock of what we can and cannot learn from mean comparisons. First, we can conclude that an unsystematic MCAR process is not plausible for the linear regression analysis—it may be reasonable for a different analysis with a different configuration of variables, but not for this model. Second,

TABLE 1.2. Standardized Mean Differences Comparing Observed and Missing Cases on Six Auxiliary Variables

	Missing data indicators		
Auxiliary variable	Pain	Pain interference	Depression
Age	0.03	–0.07	0.00
Exercise Frequency	0.30	–0.16	–0.13
Anxiety	0.43	0.11	0.33
Stress	0.24	0.00	–0.08
Control	0.07	–0.06	0.09
Disability	–0.14	–0.10	–0.01

univariate mean differences do not condition on the focal variables, so the effect sizes in Table 1.2 does not say whether a given auxiliary variable predicts missingness above and beyond the variables already in the analysis. Finally, mean differences alone do not signal a problem, as a bias-inducing MNAR-by-omission process also requires salient semipartial correlations with the analysis variables.

Identifying Correlates of Incomplete Variables

Next, I used Raykov and West's (2015) latent variable model to estimate the semipartial correlations between the auxiliary variables and the three analysis variables (the same analysis could be performed in standard statistical software using pairwise deletion). Table 1.3 gives the semipartial correlations and their significance tests. As suggested previously, semipartial correlations in the neighborhood of Cohen's (1988) medium effect size benchmark of ±0.30 are good candidates for auxiliary variables, as ignoring such variables could create a bias-inducing MNAR-by-omission process if the missing data rates are large enough (Collins et al., 2001). This rule of thumb selects three variables: anxiety, stress, and perceived control over pain. Following Collins et al.'s typology, the anxiety scale is a "category A" auxiliary variable, because it predicts missingness and uniquely correlates with depression scores. Stress and perceived control over pain can be considered "category B" variables, because they correlate with the analysis variables but do not predict their missingness. Note that these classifications are not perfect, because the patterns of correlations differ across variables (e.g., anxiety is a "category C" variable for the severe pain dummy code, because it predicts missingness but does not uniquely correlate with pain severity ratings).

Considered as a whole, the analysis results in this section offer a simple prescription: Estimate the regression model in a way that conditions on three extraneous variables that would not have appeared in the analysis had the data been complete. Doing so makes the conditionally MAR process more plausible and could improve power. Selecting additional variables based on their semipartial correlations could identify more variables than are necessary, because these bivariate associations ignore collinearity among candidate auxiliary variables. With few exceptions (e.g., an excessively large number

TABLE 1.3. Semipartial Correlations between Analysis Variables and Six Candidate Auxiliary Variables

Variable	Est.	*SE*	*z*	*p*
	Depression|Interference and Severe Pain			
Age	–.20	.06	–3.31	< .001
Exercise Frequency	–.12	.05	–2.35	.02
Anxiety	.52	.05	10.84	< .001
Stress	.46	.05	9.44	< .001
Control	–.23	.05	–4.27	< .001
Disability	.32	.06	5.61	< .001
	Interference|Depression and Severe Pain			
Age	.04	.06	0.73	.47
Exercise Frequency	–.20	.06	–3.59	< .001
Anxiety	.06	.05	1.16	.25
Stress	.02	.05	0.31	.75
Control	–.30	.05	–5.64	< .001
Disability	.12	.06	2.00	.05
	Severe Pain|Depression and Interference			
Age	.05	.06	0.86	.39
Exercise Frequency	–.13	.05	–2.53	.01
Anxiety	–.05	.06	–0.85	.39
Stress	.04	.06	0.73	.46
Control	.01	.05	0.14	.89
Disability	.09	.06	1.59	.11

of auxiliary variables, a peculiar pattern of associations; Hardt, Herke, & Leonhart, 2012; Thoemmes & Rose, 2014), there is usually no harm in casting a broad net and being overly inclusive, but you may need to restrict the size of the auxiliary set if the number of candidate variables is very large (as mentioned previously, some methods for introducing auxiliary variables are prone to convergence failures). My own experience suggests the payoff for adopting an inclusive analysis strategy is somewhat variable; leveraging additional variables sometimes affects noticeable changes in the estimates and standard errors, and other times it doesn't.

1.7 OLDER MISSING DATA METHODS

I've repeatedly referenced the analytic trio that forms the basis of this book: maximum likelihood, Bayesian estimation, and multiple imputation. These methods have been the "state of the art" for some time (Schafer & Graham, 2002), because they are capable of producing valid estimates and inferences in a wide range of applications. The literature describes numerous other approaches to missing data problems, some of which have enjoyed widespread use, while others are now little more than a historical footnote. This section describes a small collection of strategies you may still encounter in published research articles or statistical software packages: listwise and pairwise deletion, arithmetic mean imputation, regression imputation, stochastic regression imputation, and last observation carried forward imputation. These methods deal with missing data either by removing cases or by filling in the missing values with a single set of replacement scores (a process known as **single imputation**). Except for stochastic regression imputation, these methods are potentially problematic, because they invoke restrictive assumptions about the missing data process or introduce bias regardless of mechanism. In contrast, stochastic regression imputation gives valid estimates with a conditionally MAR process, but it inappropriately shrinks standard errors. I return to the artificial data in Figure 1.5 to illustrate these older approaches. To refresh, the scatterplot depicts a conditionally MAR process where participants with low perceived control over their pain were more likely to have missing depression scores.

Listwise and Pairwise Deletion

Listwise deletion (also known as **complete-case analysis**) discards the data for any case that has one or more missing values. The primary benefit of this approach is convenience, as restricting analyses to the complete cases eliminates the need for specialized software. In contrast, **pairwise deletion** (also known as **available-case analysis**) mitigates the loss of data by eliminating data records on an analysis-by-analysis basis; a prototypical example is a correlation matrix with each of its elements estimated from different subsample of cases. Reviews of published research articles suggest that deletion methods are quite common (Bodner, 2006; Jeličić, Phelps, & Lerner, 2009; Peugh & Enders, 2004; Wood, White, & Thompson, 2004), despite being characterized as being "among the worst methods available for practical applications" (Wilkinson & Task Force on Statistical Inference, 1999, p. 598).

Deletion methods have two important shortcomings: They reduce power and require an unsystematic MCAR mechanism where missingness is unrelated to the data. To illustrate the impact of the missing data process, reconsider the artificial data in Figure 1.5. The black crosshairs denote partial data records with perceived control scores but no depression values. Figure 1.9 shows the scatterplot after from removing the observations with missing depression scores. The gray contour rings convey the perspective of a drone hovering over the peak of the bivariate normal population data. As you can see, the complete score pairs are not dispersed throughout the entire range of the contour rings, and the data overrepresent the lower right quadrant of population distribution and underrepresent the upper left quadrant. As a result, the mean of the complete cases (the black dot at the center of the data) is too high along the horizontal axis (perceived control over pain) and too low along the vertical axis (depression). Not surprisingly, the systematic absence of scores from one area of the contour plot also restricts variation and distorts measures of association.

While the literature generally derides deletion methods, there are a few situations where a complete-case analysis is ideal. One such scenario occurs with linear regression models where missing values are relegated to the outcome and missingness is due to the

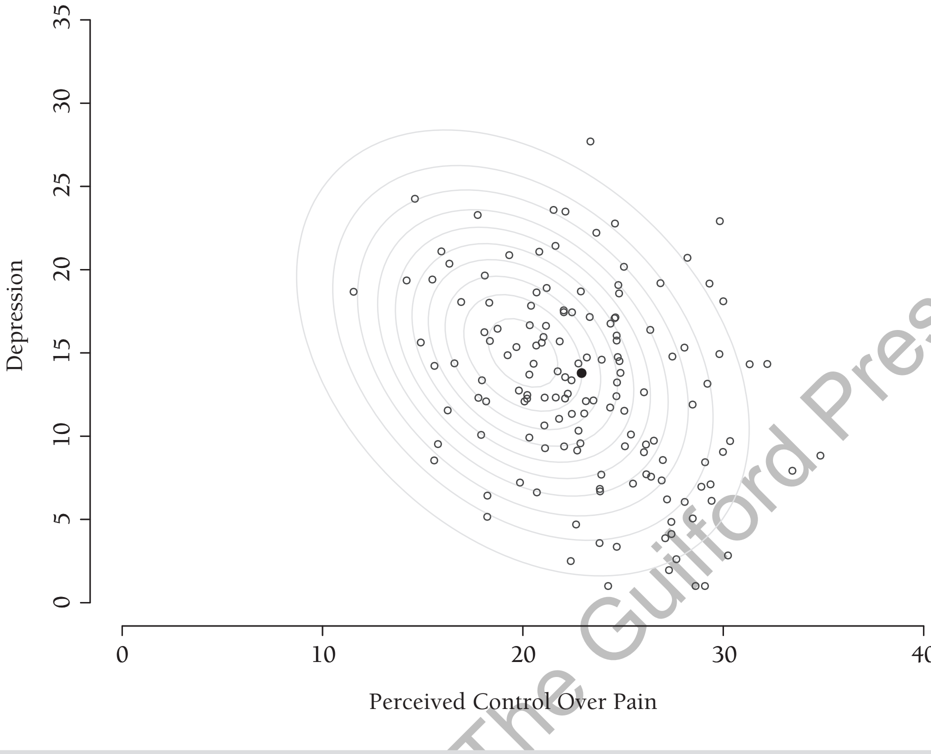

FIGURE 1.9. Scatterplot showing data points that remain after applying listwise deletion to an MAR process where 50% of the depression scores are missing for participants with lower perceived control over pain. The black circle denotes the means of the complete observations.

predictors, in which case deleting incomplete data records gives the optimal maximum likelihood estimates (Glynn & Laird, 1986; Little, 1992; von Hippel, 2007). The situation is more complicated with incomplete predictors, but deletion generally works well if missingness is unrelated to the dependent variables. This includes an MAR process where a covariate is missing as a function of another predictor, as well as an MNAR mechanism where missingness is related to the would-be values of a covariate (White & Carlin, 2010). A complete-case analysis can also provide optimal estimates of logistic regression slope coefficients in a more limited number of scenarios (Vach, 1994; van Buuren, 2012, p. 48).

Arithmetic Mean Imputation

Arithmetic mean imputation (also known as **mean substitution**) is a single imputation approach that fills in a variable's missing values with the average of its complete scores. This method has no theoretical justification and distorts parameter estimates under any missing data process. To illustrate why this is the case, Figure 1.10 shows the scatterplot of the artificial data after filling in the missing depression scores with an average of the observed scores. The gray circles in the plot represent the complete data, and the black crosshairs along a horizontal line denote score pairs with imputed data. Mean

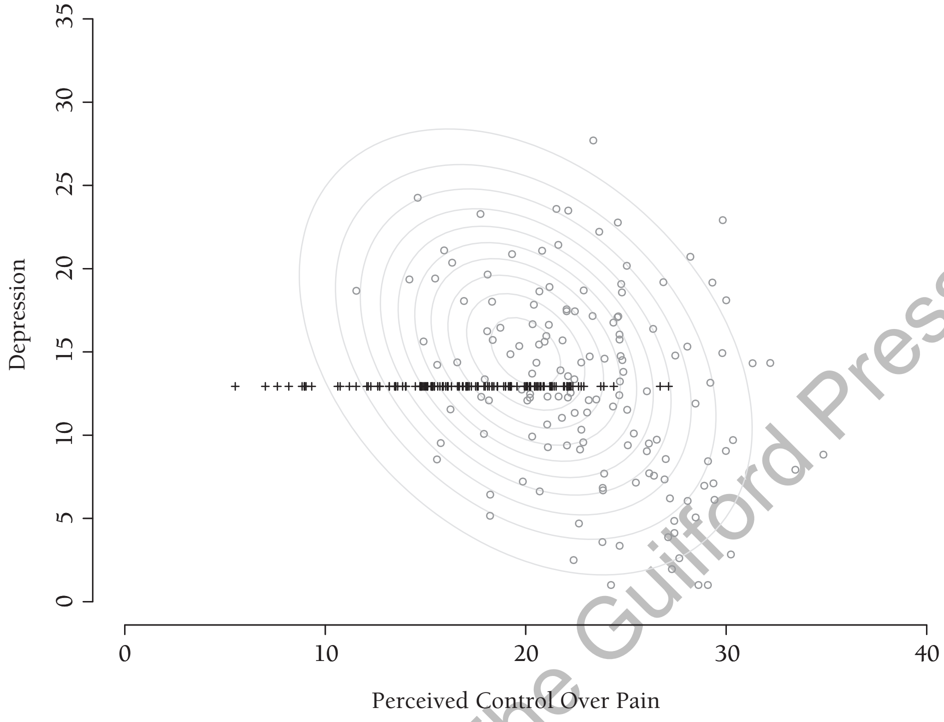

FIGURE 1.10. Scatterplot showing the data that result from applying arithmetic mean imputation to an MAR process where 50% of the depression scores are missing for participants with lower perceived control over pain. The black crosshairs denote data records with perceived control scores and imputed depression values.

imputation recoups the full set of perceived control scores, but it does a terrible job of preserving the depression distribution. As you might expect, imputing missing scores with values at the center of the distribution artificially reduces variability and attenuates measures of association (mathematically, each missing value contributes a zero to the sum of squares and sum of cross-products terms). If you focus on just the imputed score pairs, you'll notice that their correlation necessarily equals 0, because depression scores are constant. As such, you can think of mean imputation as filling in the data with scores that have no variation and no correlation with other variables. If you were going to be stranded on a desert island with only one missing data-handling procedure in your analytic suitcase, this is not the one you'd choose for your 3-hour tour.

A popular variation of mean imputation appears with questionnaire data where multiple items tap into different aspects of the same construct. For example, the continuous depression scores in the previous scatterplots result from summing item responses measuring sadness, lack of motivation, sleep difficulties, feelings of low self-worth, and so on. A common way to deal with item-level missing data is to compute a **prorated scale score** that averages the available item responses. For example, if a participant answered four out of six depression items, the prorated scale score would be the average of just four responses. The missing data literature often describes this procedure as

person mean imputation, because it is equivalent to imputing missing item responses with the average of each participant's observed scores (Huisman, 2000; Peyre, Leplege, & Coste, 2011; Roth, Switzer, & Switzer, 1999; Sijtsma & van der Ark, 2003). Like its between-person counterpart, within-person mean imputation has serious limitations that should deter researchers from using it. In particular, the method assumes an unsystematic missingness process and requires that all intrascale means and correlations are the same (Graham, 2009; Mazza, Enders, & Ruehlman, 2015; Schafer & Graham, 2002).

Regression Imputation

Regression imputation (also known as **conditional mean imputation**) replaces missing values with predicted scores from a regression equation. Regression imputation has a long history that dates back more than 60 years (Buck, 1960), and the basic idea is intuitively appealing: Variables tend to be correlated, so replacing missing values with predicted scores borrows important information from the observed data. Although this idea makes good sense, the resulting imputations can introduce substantial bias. The nature and magnitude of these biases depend on the missing data mechanism and vary across different estimands.

Regression imputation requires regression models that predict the incomplete variables from the complete variables. A complete-case analysis can generate the necessary estimates, as can maximum likelihood estimation (e.g., so-called "EM imputation"; von Hippel, 2004). Returning to the artificial data in Figure 1.5, imputation requires the regression of depression on perceived control. The following equation generates the predicted scores that serve as imputations:

$$DEPRESS_{i(\text{mis})} = \hat{\gamma}_0 + \hat{\gamma}_1 \left(CONTROL_i \right) \tag{1.13}$$

I use the γ symbol throughout the book to reference coefficients that are not part of the focal analysis, and the γ's in this equation are meant to emphasize that the regression model is a device for imputing the data. The focal analysis could be something entirely different (e.g., a correlation; the regression of perceived control on depression). The logic of regression imputation is largely the same with multivariate data, but the procedure is more cumbersome to implement, because each missing data pattern requires its own regression equation.

Figure 1.11 shows the scatterplot of the artificial data after filling in the missing depression scores with predicted values, with gray circles again representing the complete cases and black crosshairs denoting score pairs with imputed data. As you can see, the procedure recoups the full data set, but it does a subpar job of preserving the depression distribution. In particular, the imputed values lack variation, because they fall directly on the regression line. This feature also implies that the imputed score pairs have a correlation equal to 1. In effect, regression imputation suffers from the opposite problem as mean imputation, because it replaces missing values with perfectly correlated scores.

As mentioned previously, a complete-case analysis or maximum likelihood estimation can generate the coefficients for regression imputation. The latter option warrants a brief discussion, because it often confuses researchers into thinking they are applying a more sophisticated procedure than they are. This so-called "EM imputation" procedure

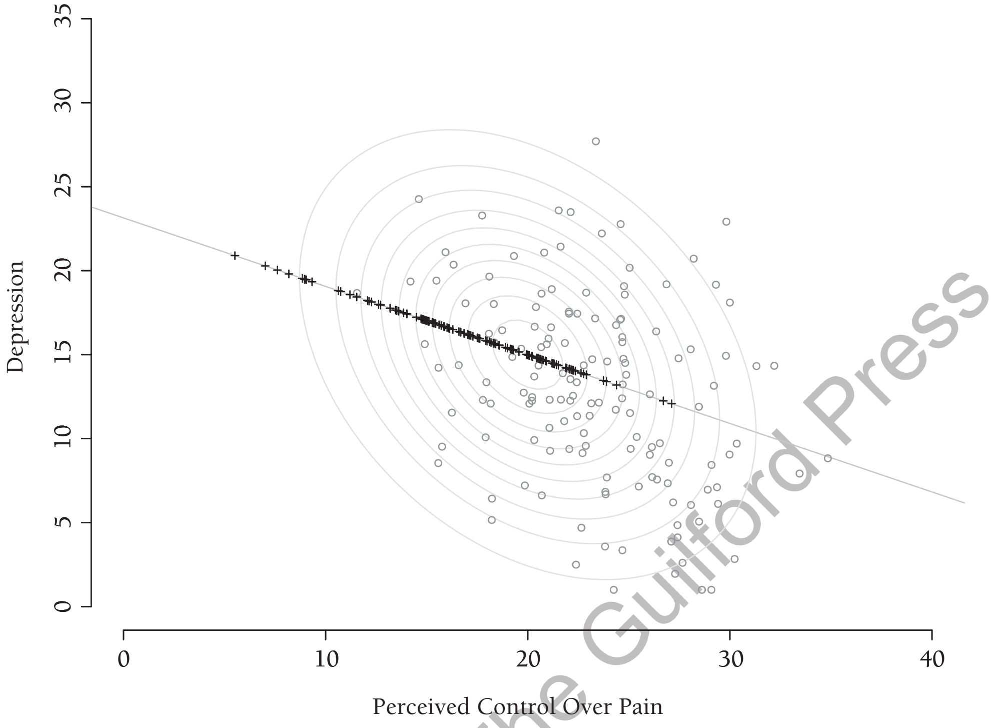

FIGURE 1.11. Scatterplot showing the data that result from applying regression imputation to an MAR process where 50% of the depression scores are missing for participants with lower perceived control over pain. The black crosshairs denote data records with perceived control scores and imputed depression values.

first uses maximum likelihood estimation (via the expectation maximization, or EM algorithm) to estimate the mean vector and covariance matrix. So far, so good, as these estimates are accurate if scores are conditionally MAR. The problem arises in the next step, where the procedure uses elements in $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ to construct regression equations that replace the missing observations with predicted values like those in Figure 1.11. Researchers sometimes characterize this method as maximum likelihood estimation when all they are really doing is using maximum likelihood to get an accurate regression equation with which to destroy the data. Interested readers can consult von Hippel (2004) for a thorough take-down of this approach, which is available in the SPSS Missing Values Analysis module, among others.

Stochastic Regression Imputation

Stochastic regression imputation also uses regression equations to predict incomplete variables from complete variables, but it takes the additional step of augmenting each predicted score with a random noise term from a normal distribution. Adding these residuals to the predicted values restores lost variability to the data and effectively eliminates the biases associated with standard regression imputation schemes. In fact, stochastic regression imputation is the only procedure in this section that is generally capable of producing unbiased parameter estimates when scores are conditionally MAR. As you will see later in the book, the core idea behind stochastic regression imputation—*an imputation equals predicted value plus noise*—resurfaces with Bayesian estimation and multiple imputation. These procedures use iterative algorithms to generate imputations over many alternate estimates of regression model parameters, but they are fundamentally sophisticated relatives of stochastic regression imputation.

Applying stochastic regression imputation to the bivariate data in Figure 1.6 again requires the regression of depression on perceived control. The residual variance from this regression plays an important role, because it defines the spread of the random noise terms. As before, substituting a participant's observed data into the right side of a regression equation gives the predicted value of the missing data point. Next, Monte Carlo computer simulation creates a synthetic residual term by drawing a random number from a normal distribution with a mean equal to 0 and spread equal to the residual variance estimate. Each imputation is then the sum of a predicted value and random noise term.

$$\begin{gathered} DEPRESS_{i(\text{mis})} = \hat{\gamma}_0 + \hat{\gamma}_1\left(CONTROL_i\right) + \dot{\varepsilon}_i \\ \dot{\varepsilon}_i \sim N_1\left(0, \hat{\sigma}_\varepsilon^2\right) \end{gathered} \tag{1.14}$$

The bottom row of the expression says that residuals are sampled from a univariate normal curve, and the dot accent on $\dot{\varepsilon}_i$ indicates that this is a synthetic value created by Monte Carlo computer simulation.

I previously introduced the possibility of drawing replacement scores from a normal curve, and Figure 1.6 shows the distribution of plausible imputations at three values of perceived control over pain. Candidate imputations fall exactly on the vertical hashmarks, but I added horizontal jitter to emphasize that more scores are located at higher contours near the regression line. Randomly selecting one of the circles from each distribution would generate an imputed depression score (technically, imputations are not restricted to the circles displayed in the graph and could be selected from anywhere in the normal distribution).

Figure 1.12 shows the scatterplot of the artificial data after filling in the missing depression scores with stochastic regression imputes. As before, the gray contour rings convey the location and elevation of the bivariate normal population distribution. Unlike the other approaches in this section, stochastic regression imputation disperses imputations throughout the entire contour plot and doesn't over- or underrepresent certain areas of the distribution. Comparing the plot to the hypothetically complete data set in Figure 1.5, the filled-in values look like good surrogates, because they preserve the center and spread of the depression scores, as well as their correlation with perceived control over pain. Although analyzing a stochastically imputed data set can provide accurate parameter estimates if values are MAR, doing so artificially shrinks standard errors and distorts significance tests; statistical software applications incorrectly treat imputes as real data when computing measures of uncertainty, such that standard errors reflect the hypothetical sampling variation that would have resulted had the data been complete. Pairing stochastic regression imputation with bootstrap resampling (Efron,

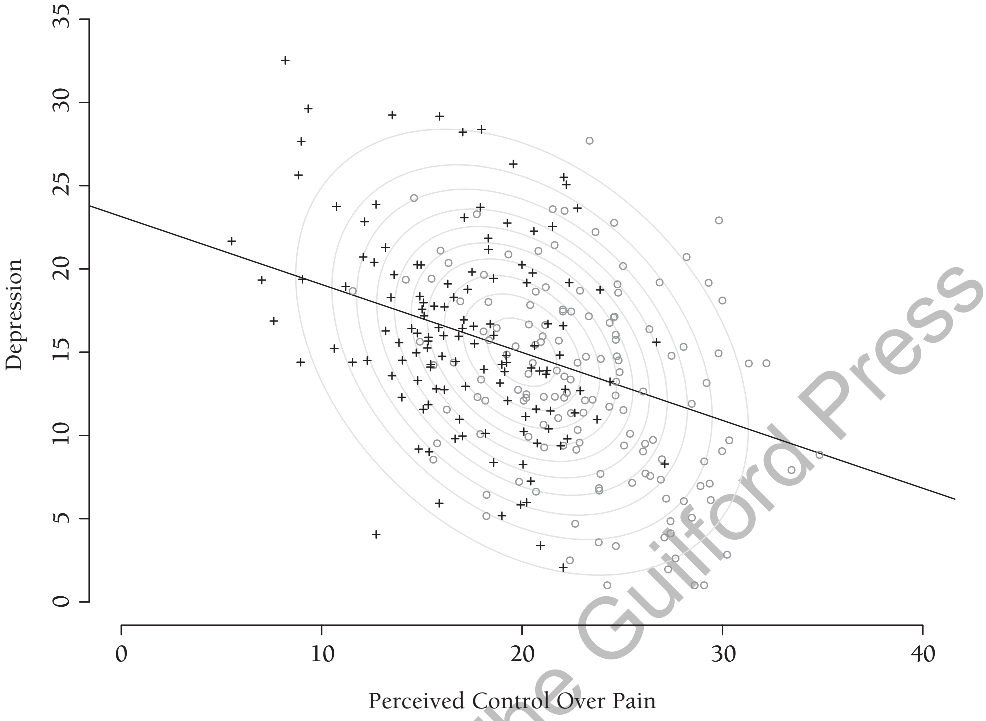

FIGURE 1.12. Scatterplot showing the data that result from applying stochastic regression imputation to an MAR process where 50% of the depression scores are missing for participants with lower perceived control over pain. The black crosshairs denote data records with perceived control scores and imputed depression values.

1987; Efron & Gong, 1983; Efron & Tibshirani, 1993) is one option for estimating measures of uncertainty (see Chapter 2) and generating and analyzing multiple sets of imputations is another (see Chapter 7).

Last Observation Carried Forward

Last observation carried forward is a missing data technique for longitudinal designs with incomplete repeated measurements. The procedure is relatively rare in the behavioral and the social sciences, and is more common in medical studies and clinical trials (Wood et al., 2004). As its name implies, last observation carried forward imputes repeated measurements with scores from the prior measurement occasion. For example, if a participant drops out after the fifth week of an 8-week study, the fifth week's score replaces all subsequent observations. To illustrate, Table 1.4 shows four waves of hypothetical depression scores for five participants, with imputed scores shown in bold typeface. As you can see, the prior measurement occasions "carry forward" regardless of whether a participant permanently attrits (e.g., the first and third data records) or has intermittent missing values (e.g., the fourth data record).

TABLE 1.4. Imputed Data from Last Observation Carried Forward

ID	Wave 1	Wave 2	Wave 3	Wave 4
		Observed data		
1	25	28	—	—
2	22	21	24	26
3	18	—	—	—
4	30	—	31	34
5	20	20	22	21
		Imputed data		
1	25	28	**28**	**28**
2	22	21	24	26
3	18	**18**	**18**	**18**
4	30	**30**	31	34
5	20	20	22	21

Last observation carried forward effectively assumes no change after the final observation or during the intermittent period where scores are missing. The conventional wisdom is that imputing the data with stable scores yields a conservative estimate of treatment group differences at the end of a study. However, empirical research shows that this isn't necessarily true, as the method can also exaggerate group differences (Cook, Zeng, & Yi, 2004; Liu & Gould, 2002; Mallinckrodt, Clark, & David, 2001; Molenberghs et al., 2004). The direction and magnitude of the bias depend on specific characteristics of the data, but the approach is likely to produce distorted parameter estimates, even with an unsystematic missingness process (Molenberghs et al., 2004). Suffice to say, there are much better strategies for dealing with longitudinal missing data.

1.8 COMPARING MISSING DATA METHODS VIA SIMULATION

The previous scatterplots suggest that older missing data methods can misrepresent distributions in ways that almost certainly introduce bias. Monte Carlo computer simulations can reveal how the tendencies depicted in the graphs unfold over many different samples and across different estimands. To this end, I used a series of simulation studies to compare listwise deletion, arithmetic mean imputation, regression imputation, and stochastic regression imputation to a "gold standard" maximum likelihood estimator for missing data. As mentioned previously, maximum likelihood missing data handling leverages the entire sample's observed data without discarding any information. The other "gold standards," Bayesian estimation and multiple imputation, are equivalent in this case (Collins et al., 2001; Schafer, 2003).

The first step of a computer simulation is to specify a set of hypothetical parameter values. Recycling the parameters that created the artificial depression and perceived control over pain data in the previous scatterplots helps visualize the procedure. Returning to Figure 1.2, the contour rings convey the perspective of a drone hovering over the peak of the bivariate normal population distribution, and the gray circles are an artificial sample of hypothetically complete data. The next step generates many artificial data sets from the population. Researchers often ask whether contemporary approaches like maximum likelihood can be used with small samples or large amounts of missing data. To examine this issue, I programmed a simulation that created 1,000 random samples of *N* = 100 from the bivariate normal population, and I deleted 50% of the artificial depression scores following one of the missing data mechanisms. The missing at completely at random process mimicked Figure 1.4, the conditionally MAR mechanism followed Figure 1.5, and the MNAR process mirrored Figure 1.7. After deleting scores, I used different missing data-handling methods to estimate three sets of parameters: the mean vector and variance–covariance matrix, coefficients from the regression of *Y* on *X* (e.g., perceived control over pain predicting depression), and coefficients from the regression of *X* on *Y* (e.g., depression predicting perceived control over pain). Any discrepancy between the average estimates and their true values reflects systematic nonresponse bias.

Missing Completely at Random

The first simulation modeled a missing (always) completely at random mechanism where missingness on *Y* (e.g., depression) was independent of the data. Table 1.5 shows the average parameter estimates for each method along with their true values. The estimates in bold typeface differ from their true values by more than 10%. Missing data

TABLE 1.5. Average Parameter Estimates from the MCAR Computer Simulation

Parameter	True value	LWD	AMI	RI	SRI	FIML
			Means, variances, covariances			
μ_X	20.00	20.02	20.02	20.02	20.02	20.02
μ_Y	15.00	14.98	14.98	14.99	15.01	14.99
σ_X^2	25.00	24.86	24.91	24.91	24.91	24.66
$\sigma_{X \cdot Y}$	–12.65	–12.83	**–6.33**	–12.87	–12.88	–12.74
σ_Y^2	40.00	40.44	**19.95**	**23.75**	40.40	39.75
$\rho_{X \cdot Y}$	–.40	–.40	**–.28**	**–.52**	–.40	–.40
			Regression of *Y* on *X*			
β_0	25.12	25.31	**20.06**	25.31	25.36	25.31
β_1 (*X*)	–0.51	–0.52	**–0.25**	–0.52	–0.52	–0.52
σ_ε^2	33.60	33.75	**18.30**	**16.48**	33.12	32.39
			Regression of *X* on *Y*			
γ_0	24.74	24.76	24.76	**28.06**	24.80	24.82
γ_1 (*Y*)	–0.32	–0.32	–0.32	**–0.54**	–0.32	–0.32
σ_r^2	21.00	20.77	22.92	**17.69**	20.56	20.19

Note. LWD, listwise deletion; AMI, arithmetic mean imputation; RI, regression imputation; SRI, stochastic regression imputation; FIML, full-information maximum likelihood.

theory predicts that listwise deletion, stochastic regression imputation, and maximum likelihood estimation are unbiased in large samples. The simulation bears this out, as the average estimates are effectively identical to the true population parameters, even with a small sample size and 50% missing data. As you might expect, mean imputation and regression imputation were prone to substantial biases. To illustrate, the solid curve in Figure 1.13 shows the sampling distribution of the correlation estimates for regression imputation, and the dashed curve shows the corresponding distribution for mean imputation. Neither method did a good job of recovering the population correlation, as the true value (the vertical line) was in the tails of both distributions. Although the presence and magnitude of the biases varied across estimands, the simulation results provide no support for these approaches on balance.

Although deletion appears to be just as good as maximum likelihood, leveraging the full sample's observed data generates estimates that are more precise, with less variation across samples. The precision difference is dramatic for some estimands and modest for others. To illustrate, the solid curve in Figure 1.14 is a kernel density plot displaying the sampling distribution of the maximum likelihood mean estimates, and the dashed curve shows the corresponding distribution for listwise deletion. As you can see, both distributions are centered at the true value of 20, but the maximum likelihood estimates are substantially closer to the truth, on average (e.g., the peak of the solid curve is higher at the true value and its tails are less thick). As a second example, Figure 1.15 shows the sampling distributions of the covariance. Maximum likelihood is again more precise, but the difference is quite modest.

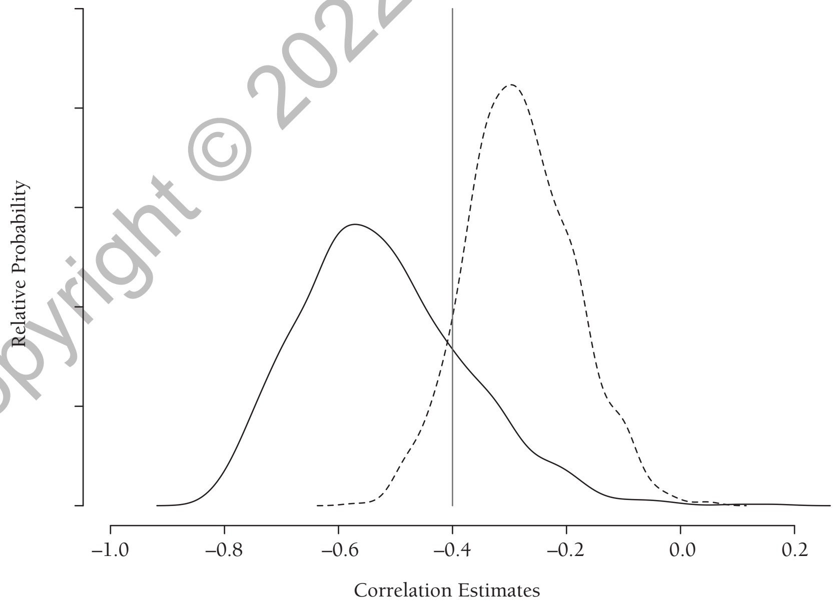

FIGURE 1.13. Kernel density plots of the correlation estimates from the MCAR computer simulation. The solid curve shows the sampling distribution of the regression imputation estimates, and the dashed curve shows the corresponding mean imputation estimates. Neither distribution is centered at the true value of –.40, indicating substantial nonresponse bias.

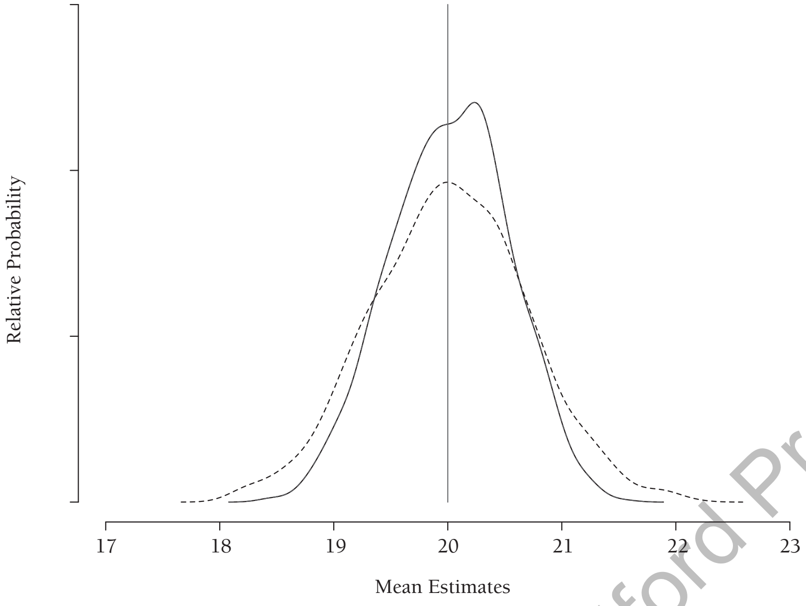

FIGURE 1.14. Kernel density plots of the *X* mean estimates from the MCAR computer simulation. The solid curve shows the sampling distribution of the maximum likelihood estimates, and the dashed curve shows the corresponding deletion estimates. Both distributions are centered at the true value of 20, but the maximum likelihood estimates are substantially closer to the true value, on average.

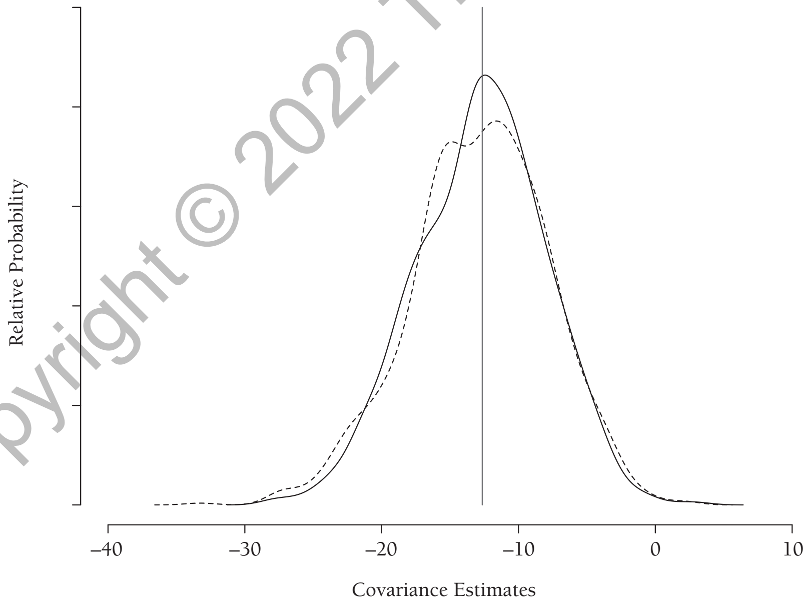

FIGURE 1.15. Kernel density plots of the covariance estimates from the MCAR computer simulation. The solid curve shows the sampling distribution of the maximum likelihood estimates, and the dashed curve shows the corresponding deletion estimates. Both distributions are centered at the true value of –12.65, but the maximum likelihood estimates are slightly closer to the true value, on average.

Missing at Random

The second simulation, which mimicked Figure 1.5, modeled a missing (always) at random mechanism where the probability of a missing *Y* score increased as the value of *X* decreased (e.g., depression scores were more likely to be missing for participants with low perceived control over pain). Table 1.6 shows the average parameter estimates for each method, along with their true values. Following the first simulation, mean imputation and regression imputation estimates were prone to bias, and the results offer no support for these procedures. A systematic missingness process was generally detrimental to the listwise deletion estimates as well. The notable exception was the regression of *Y* on *X*, where complete-case analysis gives optimal estimates when missingness does not depend on the outcome variable (Glynn & Laird, 1986; Little, 1992; von Hippel, 2007; White & Carlin, 2010). Finally, missing data theory again predicts that maximum likelihood estimation and stochastic regression imputation should be unbiased in large samples, and they are virtually so here. These results are consistent with published simulation studies showing that the percentage of missing data is not a strong determinant of bias provided that presumed mechanism is correct (Madley-Dowd, Hughes, Tilling, & Heron, 2019). Finally, stochastic regression imputation gave equivalent point estimates to maximum likelihood, but its standard errors and significance tests are untrustworthy without corrective procedures like the bootstrap.

TABLE 1.6. Average Parameter Estimates from the MAR Computer Simulation

Parameter	True values	LWD	AMI	RI	SRI	FIML
		Means, variances, covariances				
μ_X	20.00	**22.51**	19.99	19.99	19.99	19.99
μ_Y	15.00	13.74	13.74	15.01	15.02	15.01
σ_X^2	25.00	**18.64**	25.11	25.11	25.11	24.86
$\sigma_{X \cdot Y}$	–12.65	**–9.42**	**–4.66**	–12.65	–12.66	–12.52
σ_Y^2	40.00	38.44	**19.03**	**23.62**	40.05	39.57
$\rho_{X \cdot Y}$	–.40	**–.35**	**–.21**	**–.51**	–.40	–.40
		Regression of *Y* on *X*				
β_0	25.12	25.09	**17.46**	25.09	25.10	25.09
β_1 (*X*)	–0.51	–0.50	**–0.19**	–0.50	–0.50	–0.50
σ_ε^2	33.60	33.76	**18.20**	**16.54**	32.98	32.40
		Regression of *X* on *Y*				
γ_0	24.74	25.89	23.37	**27.99**	24.79	24.77
γ_1 (*X*)	–0.32	**–0.25**	**–0.25**	**–0.53**	–0.32	–0.32
σ_r^2	21.00	**16.32**	**24.04**	**18.03**	20.79	20.46

Note. LWD, listwise deletion; AMI, arithmetic mean imputation; RI, regression imputation; SRI, stochastic regression imputation; FIML, full-information maximum likelihood.

Missing Not at Random

The final simulation, which mirrored Figure 1.7, modeled a missing (always) not at random mechanism where the probability of a missing *Y* score increased as the value of *Y* itself increased (e.g., depression scores were more likely to be missing for participants with high levels of depression). Table 1.7 shows the average parameter estimates for each method, along with their true values. As you can see, all methods produced biased estimates of one or more estimands. Consistent with the MAR simulation, deletion gave accurate estimates of the regression of *X* on *Y*, because missingness did not depend on the outcome (White & Carlin, 2010). Maximum likelihood and stochastic regression imputation estimates were similarly accurate for that model but exhibited predictable biases in other analyses. Conditioning on auxiliary variables could improve the situation a little bit, but the only way to counteract nonresponse bias from a focused MNAR process like this one is to adopt a specialized analysis that introduces a nuisance model for missingness (e.g., a selection model or pattern mixture model). To illustrate one such approach, I used maximum likelihood estimation to fit a selection model that introduces an additional regression, with *Y* predicting its own missingness (the true data-generating model). The rightmost column of Table 1.7 shows that a selection model can effectively eliminate bias, but achieving that payoff requires a correctly specified nuisance model. Chapter 9 describes analysis models for MNAR processes in more detail.

TABLE 1.7. Average Parameter Estimates from the MNAR Computer Simulation

Parameter	True values	LWD	AMI	RI	SRI	FIML	FIML selection
				Means, variances, covariances			
μ_X	20.00	20.00	20.02	20.02	20.02	20.02	20.02
μ_Y	15.00	14.97	14.97	14.97	14.97	14.97	14.87
σ_X^2	25.00	24.17	25.19	25.19	25.19	24.94	24.94
$\sigma_{X \cdot Y}$	–12.65	**–9.61**	**–4.77**	**–10.06**	**–10.10**	**–9.96**	–13.17
σ_Y^2	40.00	**30.04**	**14.93**	**17.35**	**30.17**	**29.71**	42.06
$\rho_{X \cdot Y}$	–.40	**–.36**	**–.25**	**–.48**	–.36	–.36	–.39
				Regression of *Y* on *X*			
β_0	25.12	22.97	18.77	22.97	23.00	22.97	25.47
β_1 (*X*)	–0.51	**–0.40**	**–0.19**	**–0.40**	**–0.40**	**–0.40**	–0.53
σ_ε^2	33.60	**26.23**	**14.02**	**12.90**	**25.67**	**25.18**	34.64
				Regression of *X* on *Y*			
γ_0	24.74	24.80	24.82	**28.61**	25.03	25.04	24.94
γ_1 (*X*)	–0.32	–0.32	–0.32	**–0.58**	–0.34	–0.34	–0.33
σ_r^2	21.00	21.14	**23.71**	19.08	21.55	21.21	20.35

Note. LWD, listwise deletion; AMI, arithmetic mean imputation; RI, regression imputation; SRI, stochastic regression imputation; FIML, full-information maximum likelihood.

1.9 PLANNED MISSING DATA

The remainder of the chapter describes **planned missing data designs** that introduce intentional missing values as a device for reducing respondent burden or lowering research costs. The thought of intentionally creating missing values might seem odd at first, but you are probably already familiar with the idea. For example, in a randomized study with two treatment conditions, everyone has a hypothetical score from both conditions, but participants only provide a response to their assigned condition. The unobserved response to the other condition—the potential outcome or counterfactual—is missing completely at random. Viewing randomized experiments as a missing data problem is popular in the statistics literature and is a key component of Rubin's causal inference framework (Rubin, 1974; West & Thoemmes, 2010). The fractional factorial (Montgomery, 2020) is another research design that yields MCAR values. With this design, you purposefully select a subset of experimental conditions from a full factorial scheme and randomly assign participants to a restricted combination of conditions. Carefully omitting certain design cells saves resources by eliminating higher-order effects that are unlikely to be present in the data. Finally, planned missingness designs have long been a staple in educational testing applications, where examinees are administered a subset of test questions from a larger item bank (Johnson, 1992; Lord, 1962). You likely encountered a variant of this approach if you took the Graduate Record Exam.

The advent of sophisticated missing data-handling methods prompted the development of planned missingness designs that use intentional missing values to address logistical and budgetary constraints (Graham, Taylor, & Cumsille, 2001; Graham et al., 2006; Little & Rhemtulla, 2013; Raghunathan & Grizzle, 1995; Rhemtulla & Hancock, 2016; Rhemtulla & Little, 2012; Silvia, Kwapil, Walsh, & Myin-Germeys, 2014). I describe three such designs in this section: multiform designs for questionnaire data, wave missing data designs for longitudinal studies, and two-method measurement designs that pair expensive and inexpensive measures of a construct. Importantly, these designs cannot introduce bias, because they create patterns of unsystematically missing values. Of course, introducing missing data necessarily reduces power, but the loss of precision is surprisingly low in many cases.

Multiform Designs

Multiform planned missingness designs are most often associated with studies that use lengthy surveys that comprise several questionnaires and many items. Respondent burden is a major concern in these settings, because the number of items that people can reasonably answer in a single sitting is limited. A multiform design addresses this issue by administering multiple questionnaire forms that comprise different subsets of variables. For example, the classic **three-form design** (Graham et al., 1996, 2006) distributes variables into four blocks (X, A, B, and C) that are allocated across three different questionnaire forms. Each form includes the X set and is missing the A, B, or C set. Table 1.8 shows the distribution of four blocks across the three forms, with O's denoting observations and M's indicating missing values, and Figure 1.1d shows a graphical schematic of the design. Supposing that each variable set contains 25 questionnaire items, then

TABLE 1.8. Three-Form Design

	Variable set			
Form	X	A	B	C
1	O	**M**	O	O
2	O	O	**M**	O
3	O	O	O	**M**

Note. O, observed; M, missing.

survey length is reduced by 25% and participants respond to 75 rather than 100 questions. Multiform designs readily extend to include additional variable sets as needed. For example, Table 1.9 shows a six-form design from Rhemtulla and Little (2012) where respondents provide data on three out of five blocks, and Raghunathan and Grizzle (1995) and Graham et al. (2006) describe designs with even more forms.

The main downside to multiform designs (and planned missingness designs in general) is a reduction in statistical power. The impact of missing data on power and precision is complex and depends on the type of model and parameter being estimated (e.g., models with latent vs. manifest variables; correlations vs. regression slopes), as well as the effect sizes within and between blocks (Rhemtulla, Savalei, & Little, 2016). Looking at the percentage of observed responses for each variable or variable pair (sometimes called **covariance coverage**) provides some insight. To illustrate, Table 1.10 shows the covariance coverage rates for a three-form design with eight variables distributed equally across four blocks. The cell percentages reflect three tiers of precision. All things being equal, tests involving members of the X set (e.g., Y_1 and Y_2) have the most power, because these variables are complete. Variable pairs with 33% missing data introduce a second, lower tier of precision and power. This tier includes between-set associations involving a member of the X set (e.g., Y_1 and Y_3) and within-set associations between variables in the A, B, or C blocks (e.g., Y_3 and Y_4). Finally, the greatest reductions in power occur when testing associations between variable pairs with 66% missing data. This includes all between-set associations involving members of A, B, or C (e.g., Y_3 and Y_5).

TABLE 1.9. Six-Form Design

	Variable set				
Form	X	A	B	C	D
1	O	**M**	**M**	O	O
2	O	**M**	O	**M**	O
3	O	O	**M**	**M**	O
4	O	**M**	O	O	**M**
5	O	O	**M**	O	**M**
6	O	O	O	**M**	**M**

Note. O, observed; M, missing.

TABLE 1.10. Percentage of Responses within and between Blocks of a Three-Form Design

		X		A		B		C	
		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
X	Y_1	100%	100%						
	Y_2	100%	100%						
A	Y_3	66%	66%	66%	66%				
	Y_4	66%	66%	66%	66%				
B	Y_5	66%	66%	33%	33%	66%	66%		
	Y_6	66%	66%	33%	33%	66%	66%		
C	Y_7	66%	66%	33%	33%	33%	33%	66%	66%
	Y_8	66%	66%	33%	33%	33%	33%	66%	66%

With these percentages in mind, we can devise strategies for distributing variables to blocks in a way that mitigates rather than exacerbates the design's natural inefficiencies. First, pairs of variables with strong associations should appear in different blocks (Raghunathan & Grizzle, 1995; Rhemtulla & Little, 2012; Rhemtulla et al., 2016). This makes intuitive sense, because a large effect size introduces redundancy that offsets a lack of observations. This principle has important implications for studies that use multiple-item scales to measure complex constructs, where items from the same scale tend to have much stronger correlations than items belonging to different scales. Distributing a scale's items across different sets maximizes power (Graham et al., 1996, 2006; Rhemtulla & Hancock, 2016; Rhemtulla & Little, 2012), especially when using a latent variable model to examine associations among constructs (Rhemtulla et al., 2016).

Pairs of variables with weak associations are good candidates for the fully complete X set, because small effect sizes naturally require more data to achieve adequate power. Additionally, Graham et al. (2006) recommend assigning key outcome variables to the X set, as doing so maximizes power to test a study's main substantive hypotheses. Analytic work from Rhemtulla et al. (2016) supports this recommendation, as the strategy maximizes power to detect non-zero regression slopes. Including outcome variables in the X set also ensures that two-way interaction effects are estimable (Enders, 2010). Finally, the X set could also include potential determinants or correlates of *unplanned* missing data, as conditioning on such variables is necessary to satisfy the MAR assumption (Rhemtulla & Little, 2012). The power analyses in the next section highlight some of these principles.

Longitudinal Designs

Respondent burden and budgetary constraints can be particularly acute in longitudinal studies where researchers administer assessments repeatedly over time. Extending the logic of the three-form design, Graham et al. (2001) described a number of **wave missing data designs** where each participant provides data at a subset of measurement occasions. Table 1.11 shows one such design that features seven random subgroups, six of which have intentional missing data at one wave. Longitudinal planned missingness designs can be especially efficient relative to their complete-data counterparts. For example, applying the design in the table to the group-by-time interaction effect from a linear growth curve model, Graham and colleagues showed that power was 94% as large as that of a complete-data analysis. Other designs produce comparable power with even fewer data points. In situations where the total number of assessments is fixed (e.g., a grant budget can accommodate 1,000 assessments, each costing $100), Graham's chapter further showed that wave missing data designs can achieve higher power than a corresponding complete-data design; that is, collecting incomplete data from 300 participants can achieve higher power than collecting complete data from 250 participants.

Myriad configurations of patterns are possible with wave missing designs, not all of which are nearly as beneficial as the ones described earlier. Computer simulation studies provide details on a few possibilities (e.g., Graham et al., 2001; Mistler & Enders, 2011), and methodologists have outlined general strategies for identifying designs that maximize efficiency in a particular scenario. Wu, Jia, Rhemtulla, and Little (2016) developed a computer simulation tool for this purpose called SEEDMC (**SE**arch for **E**fficient **D**esigns using **M**onte **C**arlo Simulation). Their algorithm creates a design pool containing all possible planned missingness designs with a given number of measurement occasions, and it uses Monte Carlo computer simulations to create many artificial data sets for each member of the pool. Fitting a longitudinal model to each artificial data set and computing the sampling variation of the resulting estimates identifies designs with the highest relative efficiency (i.e., lowest possible sampling variation). More recently, Brandmaier, Ghisletta, and von Oertzen (2020) developed an analytic approach that estimates the measurement error of the individual change rates from a given configuration of measurement occasions. Their method selects the same optimal designs as Monte Carlo computer simulations, but it does so without intensive computations. I illustrate a combination of these strategies in Section 10.11.

Wave missing data designs are particularly useful for studies that examine change following an intervention or a treatment. However, many researchers are interested in

TABLE 1.11. Wave Missing Data Design for a Longitudinal Study

Group	% sample	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
1	16.7	O	O	O	O	O
2	16.7	**M**	O	O	O	O
3	16.7	O	**M**	O	O	O
4	16.7	O	O	**M**	O	O
5	16.7	O	O	O	**M**	O
6	16.7	O	O	O	O	**M**

Note. O, observed; M, missing.

TABLE 1.12. Cross-Sequential Design for a Developmental Study

Cohort	12	13	14	15	16	17
12	O	O	O	**M**	**M**	**M**
13	**M**	O	O	O	**M**	**M**
14	**M**	**M**	O	O	O	**M**
15	**M**	**M**	**M**	O	O	O

Note. O, observed; M, missing.

developmental processes that involve age-related change (e.g., the development of reading skills in early elementary school, the development of behavioral problems during the teenage years). **Cohort-sequential** (Duncan, Duncan, & Hops, 1996; Nesselroade & Baltes, 1979) or **cross-sequential designs** (Little, 2013; Little & Rhemtulla, 2013) are ideally suited for this type of research question. This design requires multiple age cohorts, each of which is followed over a fixed period. These shorter longitudinal studies combine to produce a much longer developmental span. To illustrate, Table 1.12 shows a cross-sequential design from a 3-year study with four age cohorts: 12, 13, 14, and 15. Notice that each cohort has three waves of intentional missing data (e.g., the 12-year-olds have missing data at ages 15, 16, and 17; the 13-year-olds have missing data at ages 12, 16, and 17; and so on).

The four 3-year studies combine to create a longitudinal design spanning 6 years, but you must be careful analyzing the data, because several bivariate associations are inestimable. For example, there are no data with which to estimate the correlation between scores at ages 12 and 15, 13 and 16, 14 and 17, and so on. This feature rules out popular multiple imputation procedures that array repeated measurements in columns (e.g., Schafer, 1997; van Buuren, 2007). However, you can readily use maximum likelihood or Bayesian estimation to fit growth models to the data, and multilevel imputation schemes that nest repeated measurements within individuals are another possibility (see Chapter 8).

Two-Method Measurement Designs

The **two-method measurement design** (Graham et al., 2006) was developed for situations in which a researcher has the choice between two measures of a construct, one of which is expensive and valid (i.e., a "gold standard" measure), the other of which is inexpensive but less valid. The basic idea is to collect the inexpensive measure from the entire sample and restrict the expensive measure to a random subset of participants. Graham et al. give an example from cigarette smoking research where self-reports with dubious validity are obtained from the entire sample and "gold standard" biochemical validators are collected from a smaller subsample. The two-method design could also be beneficial with brain imaging studies, where functional magnetic resonance imaging (fMRI) data are difficult and costly to obtain, but inexpensive behavioral measures are inexpensive and easy to collect from a much larger sample.

There are at least two ways to analyze data from a two-method measurement design. One approach is to cast the "gold standard" measure in the focal analysis model and use the inexpensive measure as an auxiliary variable. As a preview, Figure 1.16a shows a path diagram of the so-called **extra dependent variable model** (Graham, 2003) that features the auxiliary variable (the inexpensive measure) as an additional outcome. The idea is that the inexpensive measure transmits information to the expensive measure (and thus enhances the power) via its mutual association with the predictor and a correlated residual term (the double-headed curved arrow connecting the residuals). If the two measures can be cast as multiple indicators of the same construct, a second option is to analyze the data with a latent variable model similar to the one in Figure 1.16b. Graham et al. (2006) refer to this diagram as a **bias-reduction model**, because the correlated residual between the two inexpensive measures removes extraneous sources of

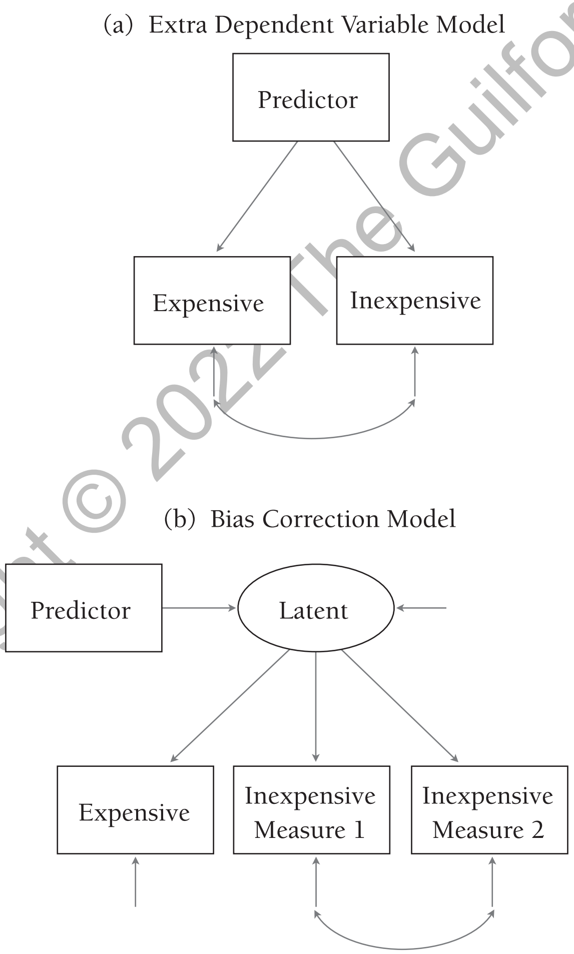

FIGURE 1.16. The top panel shows a path diagram of the extra dependent variable model, and the bottom panel is diagram of a bias-reduction model for a two-method measurement design where inexpensive and expensive measures are indicators of a latent factor.

correlation that result from invalidity, thus improving the accuracy of the structural regression coefficient connecting the covariate to the latent outcome. Graham et al. (2006) and Rhemtulla and Little (2012) provide guidelines for determining the optimal sample size ratio for the expensive measure, and Monte Carlo computer simulations are also ideally suited for this task.

1.10 POWER ANALYSES FOR PLANNED MISSINGNESS DESIGNS

This final section illustrates a power analysis for a three-form design. Section 10.10 presents a similar power study for a longitudinal growth curve model with wave missing data and unplanned missingness. I use computer simulations for this purpose, because they are relatively easy to implement and are generally applicable to virtually any analysis model. The goal of a computer simulation is to generate many artificial data sets with known population parameters and examine the distributions of the estimates across those many samples. In a power analysis, the focus shifts to significance tests, where the simulation-based power estimate is the proportion of artificial data sets that produced a significant test statistic.

The first step of a simulation is to specify hypothetical values for the population parameters. This is especially important when planning a three-form design, because the expected effect sizes dictate the assignment of variables to the four sets (e.g., variables with strong associations can be exposed to large amounts of missingness). I take a somewhat different tack that holds effect size constant to illustrate the design's natural tendencies and highlight previous findings from the literature. To this end, I considered four normally distributed variables (one variable per set) with uniformly moderate correlations equal to .30. The simulation created 5,000 random samples of $N = 250$ from this population, and I subsequently deleted data according to the three-form design in Table 1.8.

Power depends, in part, on the type of parameter being estimated (e.g., the covariance between two variables has different power than a regression slope). To illustrate this point, I fit two models to each artificial data set: a saturated model consisting of a mean vector and variance–covariance matrix, and a three-predictor linear regression model with one of the variables arbitrarily designated as the outcome. The assignment of the outcome variable to the four sets is an important consideration, so I further examined two design configurations: one with a complete predictor in the X set, and the other with a complete outcome in the X set. Figure 1.17 shows path diagrams of the four possibilities, with shaded rectangles representing blocks with missing data. I used maximum likelihood estimation to fit the analysis models to the artificial data sets, and I recorded the proportion of the 5,000 samples that produced statistically significant estimates. This proportion is a simulation-based estimate of the probability of rejecting a false null hypothesis. Maximum likelihood is the focus of the next two chapters, but for now it is sufficient to know that the estimator leverages the full sample's observed data without discarding any information. Simulation scripts are available on the companion website.

Table 1.13 gives power estimates for each correlation and regression slope along with the corresponding power values for optimal analyses with no missing data. To facilitate interpretation, the power ratios reflect complete-data power relative to that of

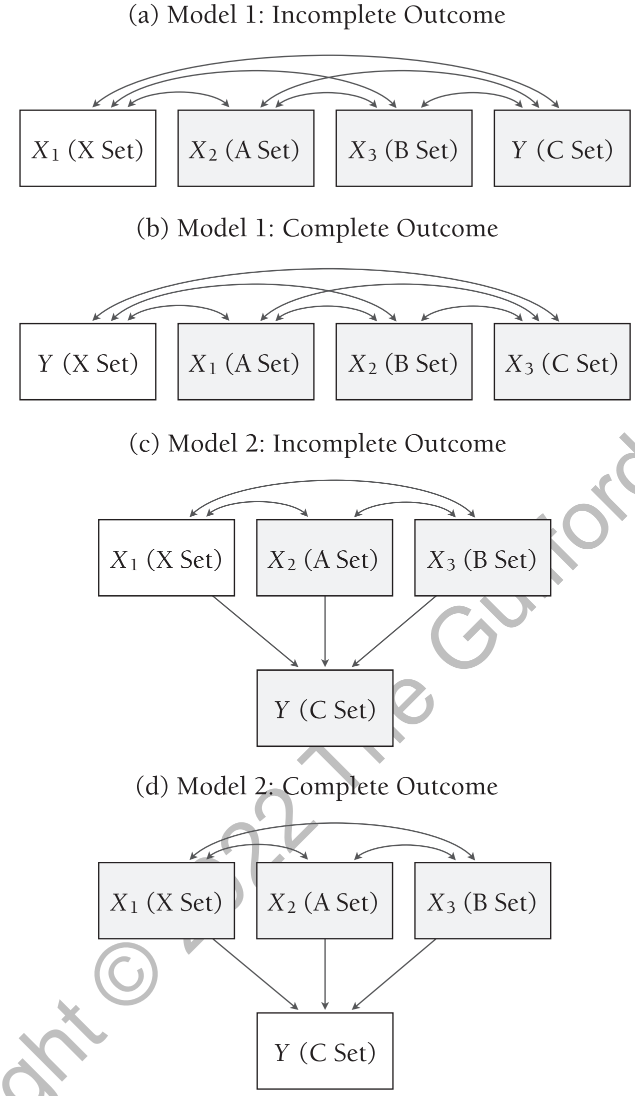

FIGURE 1.17. Path diagrams depicting two analysis models and two configurations of planned missing data. The four sets of the three-form designs are color coded, with shaded rectangles representing blocks with missing data.

a planned missingness design (e.g., 1.20 means that a complete-data analysis has 20% more power). Table 1.13 illustrates several important points, all of which echo findings from the literature. First, notice that power estimates differ by estimand, with regression slopes exhibiting lower power than correlations. This isn't necessarily surprising given that the coefficients reflect partial associations, but it nevertheless highlights the importance of considering different analyses that will be performed on the incomplete data. Second, correlations involving a complete variable in the X set (e.g., the association in the first row of the table) experienced virtually no reduction in power, even though

TABLE 1.13. Simulation-Based Power Estimates for a Three-Form Design

Parameter	Optimal power	X_1 complete		Y complete	
		Power	Power ratio	Power	Power ratio
		Correlations			
$Y \leftrightarrow X_1$	1.00	.98	1.02	.98	1.02
$Y \leftrightarrow X_2$	1.00	.83	1.21	.98	1.02
$Y \leftrightarrow X_3$	1.00	.84	1.19	.98	1.02
$X_1 \leftrightarrow X_2$	1.00	.98	1.02	.82	1.22
$X_1 \leftrightarrow X_3$	1.00	.98	1.02	.82	1.21
$X_2 \leftrightarrow X_3$	1.00	.82	1.22	.82	1.22
		Regression slopes			
$X_1 \rightarrow Y$	.85	.61	1.40	.62	1.37
$X_2 \rightarrow Y$	.86	.40	2.12	.63	1.36
$X_3 \rightarrow Y$	.86	.40	2.14	.64	1.35

33% of the other variable's scores were missing (e.g., the power advantage of a complete-data analysis was only about 2%). Third, correlations involving variable sets AB, AC, or BC (e.g., the correlation between X_2 and X_3) still had sufficient power values above .80, even though only 33% of score pairs were complete (see Table 1.10). Finally, the bottom section of Table 1.13 illustrates that assigning the outcome variable to the complete X set uniformly improves the power of all regression slopes, whereas assigning a predictor to the X set benefits only that covariate's slopes. As noted previously, assigning outcomes to the X set also ensures that all two-way interactions are estimable.

1.11 SUMMARY AND RECOMMENDED READINGS

This chapter described the theoretical underpinnings for missing data analyses, as outlined by Rubin and colleagues (Little & Rubin, 1987; Mealli & Rubin, 2016; Rubin, 1976). This work classifies missing data problems according to three different processes that link missingness to the data: an unsystematic or haphazard missing completely at random (MCAR) mechanism, a systematic conditionally missing at random (CMAR) process where missingness relates only to the observed data, and a systematic missing not at random (MNAR) mechanism where unseen score values determine missingness. From a practical perspective, these mechanisms function as statistical assumptions for a missing data analysis, and they also help us understand why not to use older methods like deletion and single imputation with a mean or predicted value.

Looking forward, most of the book is devoted to methods that naturally require a conditionally MAR assumption—maximum likelihood, Bayesian estimation, and multiple imputation. This mechanism is reasonable for many applications, and flexible software options abound. Chapter 9 describes how to modify these methods to model different MNAR processes. In the near term, maximum likelihood estimation is the next topic on the docket. Chapter 2 describes the full information estimator for complete data, and Chapter 3 applies the method to missing data problems. Finally, I recommend the following articles for readers who want additional details on topics from this chapter:

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and resrictive strategies in modern missing data procedures. *Psychological Methods, 6,* 330–351.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11,* 323–343.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology, 110,* 63–73.

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research, 151,* 53–79.

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Guilford Publications
370 Seventh Avenue
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com