Part IV

Missing Data and Dropout

17

Missing Data and Dropout: Overview of Concepts and Methods

17.1 INTRODUCTION

Missing data are a common and challenging problem in the analysis of longitudinal data. Although most longitudinal studies are designed to collect data on every individual in the sample at each time of follow-up, many studies have some missing observations. In longitudinal studies in the health sciences, missing data are the rule, not the exception. With longitudinal studies problems of missing data are far more acute than in cross-sectional studies, since non-response can occur at any occasion. An individual's response can be missing at one follow-up time and then be measured at a later follow-up time, resulting in a large number of distinct missingness patterns. Alternatively, longitudinal studies often suffer from the problem of attrition or "dropout"; that is, some individuals "drop out" or withdraw from the study before its intended completion. In either case the term "missing data" is used to indicate that an intended measurement could not be obtained.

Missing data have three important implications for longitudinal analysis. First, when data are missing, the data set is necessarily unbalanced over time since not all individuals have the same number of repeated measurements at a common set of occasions. As a result missing data create complications for methods of analysis that require balanced data. However, imbalance created by missingness is not a concern for the regression methods described in Parts II and III. Second, when there are missing data, there must necessarily be some loss of information. That is, missing data reduce the precision with which changes in the mean response over time can be estimated. Not surprisingly, the reduction in precision is directly related to the amount of missing data; that is, the greater the amount of missing data, the greater is the decrease in

Applied Longitudinal Analysis, Second Edition. By Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware.

precision. The loss of precision can also depend to some extent on the method of analysis; for example, analyses restricted to subjects with complete data will generally be less efficient than methods that use all available data. The location of the missing data (e.g., missingness spread sporadically over many subjects, or concentrated at a specific set of time points in a few subjects), and how highly correlated the missing data are with the observed data, will also affect loss of precision. Finally, under certain circumstances missing data can introduce bias and thereby lead to misleading inferences about changes in the mean response. It is this last factor, the potential for serious bias, that complicates the analysis of partially missing longitudinal data. As a result the reasons for any missing data, often referred to as the *missing data mechanism*, must be carefully considered.

This is the basis for an important theme that will be emphasized throughout this chapter: when data are missing, we must carefully consider why they are missing. Some types of missing data are relatively benign and do not complicate the analysis; others are not and can potentially introduce bias in the estimates of the regression parameters. The following two examples of partially missing longitudinal data will help illustrate this point.

The first example is from the Six Cities Study of Air Pollution and Health, discussed in Sections 8.8 and 9.6. This was a longitudinal study designed to characterize lung function growth as measured by changes in pulmonary function in children and adolescents. Most of the children were enrolled in the first or second grade (between the ages of six and seven) and measurements were obtained annually until graduation from high school or loss to follow-up. At each annual examination, spirometry, the measurement of pulmonary function, was performed. Due to late entry into the study and loss to follow-up or attrition, the number of measurements of pulmonary function of study children varied from a minimum of 1 to a maximum of 12. The major reason for late entry or attrition was moving in or out of the school district. Let us focus on this main reason for missing data. If a child changed school district because of employment relocation by her parents, then the missing data mechanism can be thought of as unrelated to the child's pulmonary function. On the other hand, if a child moved out of the school district because she developed respiratory problems (e.g., relocating to an area with either better air quality or improved access to health care), then missingness is related to the child's pulmonary function.

The second example is from the Muscatine Coronary Risk Factor (MCRF) study, introduced in Section 1.3 and analyzed in Section 13.4. This was a longitudinal survey of school-age children in Muscatine, Iowa, examining the development and persistence of risk factors for coronary disease. In the MCRF study, weight and height measurements of five cohorts of children, initially aged 5–7, 7–9, 9–11, 11–13, and 13–15 years, were obtained biennially from 1977 to 1981. On the basis of a comparison of their weight to age–gender-specific norms, children were classified as obese or not obese. The study protocol required parental consent prior to each measurement. One objective of the MCRF study was to determine whether the risk of obesity increased with age and whether patterns of change in obesity were the same for boys and girls. Although each child was eligible to participate in all three surveys, there was a substantial amount of missing data on obesity, with less than 40% of the

children providing complete data at all three measurement occasions. The two main reasons for missing data were: (1) failure to obtain consent and (2) the child's absence from school on the day of examination. Let us focus on these two reasons for missing data. Suppose that parents of children who were obese were either more or less likely to sign the consent form than parents of children who were not obese. For example, parents of children who were obese might have been more likely to sign the consent form due to concerns about the adverse health effects of obesity; conversely, they might have been less likely to sign the consent form due to concerns that participation in the study could be a source of embarrassment for their children. In either case the reason for missing data on weight and height is related to the obesity status of the child. Similarly missingness is related to the obesity status of the child if children who were obese were more likely to be absent on the day of examination (e.g., due to embarrassment about being overweight). On the other hand, suppose that a child was absent on the day of examination because of employment relocation by her parents (completely unrelated to the health of the child). Then missingness does not depend on the child's obesity status.

These two examples show that there can be more than a single cause of missing data and that reasons for missing data may or may not be related to the outcome of interest. When data are missing for reasons unrelated to the outcome of interest, the impact of missing data is relatively benign and does not complicate the analysis. When it is related to the outcome, somewhat greater care is required because there is potential for bias when individuals with missing data differ in important ways from those with complete data.

In this chapter we review three general models for missing data that differ in terms of assumptions concerning whether missingness is related to observed and unobserved responses. We also discuss the implicit assumptions about missing data that underlie the methods for longitudinal analysis described in Parts II and III. We illustrate the main distinctions between the three general models for missing data for the common problem of dropout. Finally, we provide an overview of some alternative methods for handling dropout in longitudinal studies. A more in-depth discussion and application of two important methods for handling missing data, multiple imputation and inverse probability weighted methods, can be found in Chapter 18.

17.2 HIERARCHY OF MISSING DATA MECHANISMS

To obtain valid inferences from partially missing longitudinal data, we must consider the nature of the "missing data mechanism." Ordinarily the missing data mechanism is not under the control of the investigators and often is not well understood. Instead, assumptions are made about the missing data mechanism and the validity of the analysis depends on whether these assumptions hold. When reporting the results of a longitudinal analysis, it is important to be explicit about the assumptions made regarding the reasons for missing data.

The missing data mechanism can be thought of as a probability model for the distribution of a set of response indicator variables. These response indicator variables

take the value 1 when an intended measurement of the response is obtained and the value 0 otherwise. For example, suppose that the design of the study calls for n measurements per subject. That is, we intend to take n repeated measures of the response variable on the same individual. A subject with a *complete* set of responses has an $n \times 1$ response vector denoted by

$$Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in})';$$

this is a slight abuse of the notation adopted in previous chapters where Y_i contained not the possible set of observations but the values actually observed. Because of missing data, some of the components of Y_i are not observed for at least some individuals. We let R_i be an $n \times 1$ vector of response indicators

$$R_i = (R_{i1}, R_{i2}, ..., R_{in})',$$

with $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ if Y_{ij} is missing. In addition associated with Y_i is an $n \times p$ matrix of covariates, X_i. We do not consider missingness in the covariates; that is, we assume that any time-varying covariates are fixed by the study design. Given R_i, the complete set of responses, $Y_i = (Y_{i1}, \ldots, Y_{in})'$, can be partitioned into two components Y_i^O and Y_i^M, corresponding to those responses that are observed and missing, respectively. That is, Y_i^O denotes the vector of *observed* responses on the i^{th} subject, and Y_i^M denotes the complementary set of responses that are missing. The random vector R_i is recorded for all individuals. Also, given R_i, the target population of interest can be divided or stratified into a number of distinct sub-populations defined by the missing data patterns (including the sub-population of "completers"). Thus R_i can also be thought of as a stratification variable that divides the target population into a number of sub-populations. This is illustrated in Table 17.1, where the first response, Y_1, perhaps denoting a baseline response, is fully observed, but $Y_2, \ldots, Y_n$ are missing intermittently.

A hierarchy of three different types of missing data mechanisms can be distinguished by considering how R_i is related to Y_i:

1. *Missing Completely at Random* (MCAR),
2. *Missing at Random* (MAR), and
3. *Not Missing at Random* (NMAR).

The hierarchy of missing data mechanisms is useful because the type of missing data mechanism determines the appropriateness of different methods of analyses, for example, maximum likelihood, generalized least squares (GLS), and GEE. We discuss this topic later in the chapter. However, the nomenclature is not intuitive and leads to much confusion among statisticians and practitioners alike. A major objective of this chapter is to explain these mechanisms in a more intuitive manner so that the reader gains a better appreciation for their usage.

Much of the remainder of this section is devoted to a detailed explanation, with concrete examples, of this classification of missing data mechanisms in the context of

Table 17.1 Schematic representation of R, the vector of response indicators, as a stratification variable.

Response Indicators						Response Vector[a]					
R_1	R_2	R_3	R_4	$\cdots$	R_n	Y_1	Y_2	Y_3	Y_4	$\cdots$	Y_n
1	1	1	1	$\cdots$	1	y_1	y_2	y_3	y_4	$\cdots$	y_n
1	0	1	1	$\cdots$	1	y_1	$*$	y_3	y_4	$\cdots$	y_n
1	1	0	1	$\cdots$	1	y_1	y_2	$*$	y_4	$\cdots$	y_n
1	1	1	0	$\cdots$	1	y_1	y_2	y_3	$*$	$\cdots$	y_n
.	.	.	.	$\cdots$	.	.	.	.	.	$\cdots$	.
.	.	.	.	$\cdots$	.	.	.	.	.	$\cdots$	.
.	.	.	.	$\cdots$	.	.	.	.	.	$\cdots$	.
1	0	0	0	$\cdots$	1	y_1	$*$	$*$	$*$	$\cdots$	y_n
1	0	0	0	$\cdots$	0	y_1	$*$	$*$	$*$	$\cdots$	$*$

[a]The $*$ denotes missing value.

longitudinal studies. We begin each description with the formal definition expressed as conditions on the probability distribution of the response indicators, R_i. We then provide some examples and explain the consequences of each type of missingness for the distribution of the observed data. Once the main distinctions are understood, we can describe the implicit assumptions about the missing data mechanism made by different methods for analyzing longitudinal data.

Missing Completely at Random (MCAR)

Data are said to be missing completely at random (MCAR) when the probability that responses are missing is unrelated to either the specific values that, in principle, should have been obtained or the set of observed responses. That is, longitudinal data are MCAR when R_i is independent of both Y_i^O and Y_i^M, the observed and unobserved components of Y_i, respectively. To better understand this missing data mechanism, consider the bivariate case where $Y_i = (Y_{i1}, Y_{i2})'$, Y_{i1} is assumed to be fully observed and Y_{i2} is sometimes missing. In that case we require only a single response indicator, with $R_{i2} = 1$ if Y_{i2} is observed and $R_{i2} = 0$ if Y_{i2} is missing. If Y_{i2} is MCAR, then

$$\Pr(R_{i2} = 1 | Y_{i1}, Y_{i2}, X_i) = \Pr(R_{i2} = 1 | X_i),$$

and the probability that Y_{i2} is missing does not depend on the observed value of Y_{i1} or the value of Y_{i2} that, in principle, should have been obtained. Missingness in Y_{i2} is simply the result of a chance mechanism that does not depend on observed or unobserved components of Y_i.

An example where partially missing longitudinal data are MCAR is the "rotating panel" study design. In this study design, commonly used in health surveys to reduce response burden, individuals rotate in and out of the study after providing a pre-determined number of repeated measures. However, the number and timing of the measurements is determined by design. The decision about whether to obtain a measurement on an individual at any specific occasion is made a priori by the investigators and is not related to the vector of responses; that is, R_i is unrelated to Y_i. In this example the missing data mechanism is under the control of the investigators and is well understood. Another example where missing data are MCAR is in the Six Cities Study of Air Pollution and Health, when children changed school district because of employment relocation by their parents (for reasons completely unrelated to the health of their children). Here the reason for missing data is unrelated to the children's pulmonary function.

In the definition of MCAR given above, missingness can depend on the covariates, X_i. This raises a subtle, but important, point. Under MCAR, the response vector Y_i is conditionally independent of R_i, given the covariates X_i. However, this conditional independence of Y_i and R_i may not hold when conditioning on only a subset of the covariates. This has the following important implication. When an analysis is based on a subset of X_i that excludes a covariate that is predictive of R_i, the missing data can no longer be considered MCAR. For example, in a clinical trial missingness may be related to side effects of the treatments. However, side effects is a covariate that would not ordinarily be included in the analysis model that evaluates treatment effects. If side effects is excluded from the analysis model, the missing data can no longer be considered MCAR; in Chapter 18 we discuss how this more complex case can be handled. When

$$\Pr(R_i|Y_i^O, Y_i^M, X_i) = \Pr(R_i|X_i), \tag{17.1}$$

the data are said to have *covariate-dependent* missingness and use of the term MCAR is sometimes restricted to the case where

$$\Pr(R_i|Y_i^O, Y_i^M, X_i) = \Pr(R_i). \tag{17.2}$$

In our discussion of missing data mechanisms we do not make this subtle distinction. Instead, we define MCAR using (17.1) and simply assume that X_i in (17.1) contains all relevant covariates for predicting both Y_i and R_i.

The essential feature of MCAR is that the observed data can be thought of as a random sample of the complete data. The result is that the moments (e.g., means, variances, and covariances) and, indeed, the distribution of the observed data do not differ from the corresponding moments or distribution of the complete data. Thus, "completers" can be regarded as a random sample from the target population, albeit with a smaller sample size than intended. This has important implications for the analysis of longitudinal data restricted to subjects with complete response vectors. In particular, any method of analysis that yields valid inferences in the absence of missing data will also yield valid inferences when missing data are MCAR and the analysis is restricted to the "completers." The latter is often referred to as a "complete-case"

analysis. With data MCAR it is legitimate (although possibly wasteful) to remove subjects with any missing data from the analysis since we can regard them as randomly chosen without regard to their data values. This feature of MCAR allows one to do a complete-case analysis without being concerned that the results might be biased by excluding those with missing data.

A similar result holds for subjects with some missing data. The responses actually obtained, Y_i^O, have the same distribution as the corresponding elements of the completers. As a result all available data can be used to give valid estimates of moments such as means, variances, and covariances. For example, if we modify our bivariate example to allow data to be MCAR either at time 1 or time 2, then subjects with only one observation can be used along with the complete cases to estimate means and variances; only the complete cases can here be used to estimate the covariance. In longitudinal designs with more observations per subject, the observed cases with at least two observations contribute to covariance estimation. As a result methods for longitudinal analysis that incorporate all of the available observations will yield valid inferences when missing data are MCAR. This includes all of the methods that were discussed in Parts II and III of this book.

These properties of MCAR follow directly from the definitions in (17.1) and (17.2). They can be used to show that when the missing data mechanism is MCAR, the distribution of Y_i (given X_i) is the same in each of the distinct sub-populations defined by the missing data patterns (including the sub-population of "completers" or subjects with no missing responses). It also implies that these distributions coincide with the distribution of Y_i (given X_i) in the target population of interest. Moreover, when the missing data mechanism is MCAR, the distribution of the observed components Y_i^O for any distinct sub-population defined by the missing data patterns coincides with the distribution of the same components of Y_i in the target population.

Finally, we note that with MCAR, the distribution of Y_i^M for any distinct sub-population defined by the missing data patterns coincides with the distribution of the same components of Y_i for the "completers." For example, in the bivariate case, MCAR implies that the distribution of Y_{i2} for those missing Y_{i2} is the same as the distribution of Y_{i2} for those with no missing responses.

Missing at Random (MAR)

Data are said to be missing at random (MAR) when the probability that responses are missing depends on the set of observed responses but is conditionally unrelated to the specific missing values that, in principle, should have been obtained. Specifically, longitudinal data are MAR when R_i is conditionally independent of Y_i^M, given Y_i^O,

$$\Pr(R_i|Y_i^O, Y_i^M, X_i) = \Pr(R_i|Y_i^O, X_i). \tag{17.3}$$

Let us return to the bivariate case where $Y_i = (Y_{i1}, Y_{i2})'$, Y_{i1} is fully observed, and Y_{i2} is sometimes missing. If Y_{i2} is MAR, then

$$\Pr(R_{i2} = 1|Y_{i1}, Y_{i2}, X_i) = \Pr(R_{i2} = 1|Y_{i1}, X_i),$$

and the probability that Y_{i2} is missing depends on the observed value of Y_{i1}. However, given Y_{i1}, the probability that Y_{i2} is missing does not depend on the value of Y_{i2} that should have been obtained. Put another way, if subjects are stratified on the basis of similar values for Y_{i1}, missingness in Y_{i2} within strata is simply the result of a chance mechanism that does not depend on values of Y_{i2}.

Longitudinal data that are MAR might arise when a study protocol requires that a subject be removed from the study whenever the value of an outcome variable falls outside of a certain clinical range of values. In that case missingness in Y_i is under the control of the investigator and is related to observed components of Y_i only.

Another example where missing data are MAR is in the Six Cities Study of Air Pollution and Health, when children moved out of the school district because they developed respiratory problems. If the decision to relocate could be predicted based only on the recorded history of pulmonary function measurements (i.e., the observed components of Y_i only), then the missing data are MAR. However, the MAR assumption would not hold if the decision to relocate was based on some extraneous variable, unavailable to the investigators, that was predictive of the future but unobserved, pulmonary function measurements.

Because the missing data mechanism now depends on Y_i^O, the distribution of Y_i in each of the distinct sub-populations defined by the missing data patterns is not the same as the distribution of Y_i in the target population. This has important consequences for analysis. One is that a "complete-case" analysis is not valid and can produce biased estimates of change in the mean response over time. Furthermore the distribution of Y_i^O, the observed components of Y_i, in these sub-populations does not coincide with the distribution of the same components of Y_i in the target population. Therefore the sample means, variances, and covariances based on the available data are biased estimates of the corresponding parameters in the target population. This feature of MAR will be illustrated in the context of dropout in Section 17.4.

With MAR, the observed data cannot be viewed as a random sample of the complete data, but there is an important implication for the distribution of the missing data. The distribution of each individual's missing values, Y_i^M, conditioned on the observed values, Y_i^O, is the same as the conditional distribution of the corresponding observations among the complete cases, conditional on those complete cases having the same values as Y_i^O. In other words, if we stratify on values of Y_i^O, the distribution of Y_i^M is the same as the distribution of the corresponding observations in the complete-case and target populations. As a result missing values can be validly predicted using the observed data and a model for the joint distribution. However, the validity of the predictions of the missing values rests on having correctly specified both the model for the mean and the model for the covariance (when the responses have a multivariate normal distribution). The model for the covariance must be correctly specified because conditional moments (e.g., conditional means) depend on both the mean response vector and the covariance.

For example, if the responses are assumed to have a multivariate normal distribution, then predictions of the missing values are based on the conditional mean of Y_i^M, given Y_i^O. Using well-known properties of the multivariate normal distribution, the

conditional mean of Y_i^M, given Y_i^O, can be expressed as

$$E(Y_i^M|Y_i^O) = \mu_i^M + \Sigma_i^{MO}\Sigma_i^{O^{-1}}(Y_i^O - \mu_i^O),$$

where μ_i^M and μ_i^O denote those components of the mean response vector corresponding to Y_i^M and Y_i^O, and Σ_i^O and Σ_i^{MO} denote those components of the covariance matrix corresponding to the covariance among the elements of Y_i^O and the covariance between Y_i^M and Y_i^O. The important aspect of the expression given above is the dependence of the prediction of Y_i^M on both the mean response vector

$$\mu_i = \begin{pmatrix} \mu_i^O \\ \mu_i^M \end{pmatrix},$$

and the covariance matrix

$$\Sigma_i = \begin{pmatrix} \Sigma_i^O & \Sigma_i^{OM} \\ \Sigma_i^{MO} & \Sigma_i^M \end{pmatrix}.$$

When missing data are MAR, we must correctly model the entire joint distribution of Y_i, $f(Y_i|X_i)$ (e.g., both the mean and covariance when Y_i is assumed to have a multivariate normal distribution) to obtain valid estimates of β (and Σ_i).

With MAR, the missing values can be predicted using the observed data and a model for the joint distribution of Y_i. But one does not need to use the model for $\Pr(R_i|Y_i^O, X_i)$ as a function of X_i and Y_i^O, only a model for Y_i given X_i. Since MCAR is a special case of MAR, the same is also true of MCAR; namely one does not need to use the model for $\Pr(R_i|Y_i, X_i)$ to obtain valid likelihood-based inferences, only a model for $f(Y_i|X_i)$. Notice that not using $\Pr(R_i|Y_i, X_i)$ in the analysis has the important implication that we do not need to even posit a specific model for $\Pr(R_i|Y_i, X_i)$ other than to say it does not depend on the missing observations. Since it is common to use a model for $f(Y_i|X_i)$, valid likelihood-based analyses can be obtained with MAR or MCAR data with no extra assumptions, other than the general statement of MCAR or MAR. For this reason MCAR and MAR are often referred to as *ignorable* mechanisms. A caveat concerning this use of the term *ignorable*, when data are MAR, it emphatically does not mean we can ignore the missing data problem and use any complete-case or available-data analysis we desire. Instead, the ignorability refers to the fact that once we establish that $\Pr(R_i|Y_i, X_i)$ does not depend on missing observations, we can ignore $\Pr(R_i|Y_i, X_i)$ and obtain a valid likelihood-based analysis provided that we have a correct model for $f(Y_i|X_i)$. That is, ML estimation of β in the linear models discussed in Part II is valid when data are MAR provided that the multivariate normal distribution has been correctly specified.

In contrast to a full-likelihood analysis, standard applications of generalized least squares (GLS) that only require a model for the mean response, but do not assume a multivariate normal distribution for the response vector, no longer provide valid estimates of β. That is, when data are MAR, GLS based only on the means, variances, and covariances of the available data can yield biased estimates of β. This is because

the sample means, variances, and covariances based on the available data (or based on the "completers") are biased estimates of the corresponding parameters in the target population. Moreover, with GLS, the means, variances, and covariances may possibly be misspecified. In a similar way the generalized linear mixed effects models (GLMMs) described in Part III fully specify the joint distributions of both the vector of responses and the vector of random effects. As a result conventional likelihood-based analyses of the incomplete data using GLMMs yield valid inferences when data are missing at random (MAR), provided that the likelihood has been correctly specified; this property, however, does not extend to approximate methods such as PQL. This is in contrast to the GEE methods for analyzing discrete longitudinal data described in Part III; the GEE methods require a model for the mean response but do not specify the multivariate joint distribution for the response vector. As a result standard GEE methods do not provide valid estimates of the regression parameters when data are MAR but not MCAR. However, both the GLS and GEE estimators of β can be adapted to provide a valid analysis by explicitly modeling $\Pr(R_i|Y_i, X_i)$ and weighting the analysis accordingly; the intuition for "weighted methods" is discussed in Section 17.5 and weighting methods are described in greater detail in Chapter 18.

The subtle distinction between MCAR and MAR is often not well understood. We find that statisticians and empirical researchers regularly confuse the definition of MAR with MCAR (and admittedly, the choice of terminology has not helped matters). As we will see in the next section, the distinction between MAR and MCAR has very important implications for the validity of different methods of analysis of longitudinal data. The MAR assumption is far less restrictive on $\Pr(R_i)$ than MCAR and may be considered to be a more plausible assumption about missing data in many applications. Of note, although the MAR assumption is less restrictive in the sense of restrictions on $\Pr(R_i)$, it can be considered more restrictive in terms of what methods of analyses are appropriate. In our view, the MAR assumption should be the default assumption for the analysis of partially missing longitudinal data unless there is a strong and compelling rationale to support the MCAR assumption.

Not Missing at Random

The third type of missing data mechanism is referred to as *not missing at random* (NMAR). Missing data are said to be NMAR when the probability that responses are missing is related to the specific values that should have been obtained. That is, the conditional distribution of R_i, given Y_i^O, is related to Y_i^M, and

$$\Pr(R_i|Y_i^O, Y_i^M, X_i)$$

depends on at least some elements of Y_i^M. Let us return to the bivariate case where $Y_i = (Y_{i1}, Y_{i2})'$, Y_{i1} is fully observed, and Y_{i2} is sometimes missing. If missingness in Y_{i2} is NMAR, then

$$\Pr(R_{i2} = 1|Y_{i1}, Y_{i2}, X_i)$$

depends on the potentially unobserved value of Y_{i2}. An NMAR mechanism is often referred to as *nonignorable* missingness. The term *nonignorable* refers to the fact that

the missing data mechanism cannot be ignored when the goal is to make inferences about the distribution of the complete longitudinal responses.

An example where longitudinal data are NMAR arises when the outcome variable is a measure of "quality-of-life" and subjects fail to complete the instrument or questionnaire on occasions when their quality-of-life is compromised. Another example where missing data are NMAR is in the Muscatine Coronary Risk Factor (MCRF) study, when parents of children who were obese were either more or less likely to sign the consent form than parents of children who were not obese. In that case missingness on weight and height is related to the obesity status of the child.

Sometimes the term *informative* is used to describe data that are NMAR; missingness is informative in the sense that the missingness (i.e., a component of R_i is equal to 0) informs us about the distribution of the missing observations. Specifically, the distribution of Y_i^M, conditional on Y_i^O, is not the same as that in the "completers" or in the target population, but rather, the distribution of Y_i^M depends on Y_i^O and on $\Pr(R_i|Y_i, X_i)$. Thus the model assumed for $\Pr(R_i|Y_i, X_i)$ is crucial; it must be included in the analysis, and the specific model chosen can drive the results of the analysis.

17.3 IMPLICATIONS FOR LONGITUDINAL ANALYSIS

The statistical methods for analyzing longitudinal data presented in earlier chapters of the book can accommodate incomplete data. However, valid inferences from partially missing longitudinal data require assumptions about the missing data mechanism. In this section we summarize the key assumptions about missing data required for valid inferences when applying the techniques described in Parts II and III.

When the missing data mechanism is MCAR, individuals with missing data are a random subset of the sample. In this case the observed values of the responses are a random subsample of all values of the responses, and no bias will arise with almost any method of analysis of the data (either the available data or the data on the "completers" only). In particular, all of the methods discussed in Parts II and III will yield valid estimates of mean response trends (and within-subject associations) if the missing data can be assumed to be MCAR.

When the missing data mechanism is MAR, individuals with missing data are no longer a random subset of the sample. Only when stratified on their observed outcomes (i.e., conditional on Y_i^O) can they be considered a random subset of the sample belonging to that stratum. As a result the observed values are not necessarily a random subsample of the responses. In particular, the distribution of Y_i^O, the observed components of Y_i, differs from the distribution of the same components of Y_i in the target population. This implies that based on the available observations the sample means at each occasion (and the covariances) provide biased estimates of the means (and covariances) in the target population. Similarly analyses restricted to the data from the completers also yield biased estimates of the means (and covariances). When missing data are MAR, but not MCAR, complete-case methods and standard GEE methods based on all of the available observations yield biased estimates of

mean response trends. In contrast, likelihood-based methods that correctly specify the entire joint distribution of the responses yield valid estimates when missing data are MAR. However, there is a subtle, but important, proviso: the models for both the mean response and the within-subject association must be correctly specified. Thus, when missing data are MAR, the likelihood-based methods discussed in Part II provide valid inferences about changes in the mean response over time provided that the covariance matrix has been correctly modeled. Similarly the methods discussed in Chapter 14 provide valid estimates of the fixed effects provided that the random effects structure has been correctly specified. In summary, when missing data are MAR, but not MCAR, inferences about the mean response are sensitive to any form of misspecification of the joint distribution of the vector of responses. Accordingly, if longitudinal data are incomplete, somewhat greater care must be exercised when modeling the within-subject association.

The standard GEE approach requires that we have a model for the expected value of the observations given the covariates. With MAR, this marginal model for the mean response will generally not hold for the observed data, so the validity of the analysis is compromised. Methods have been devised for making adjustments to the analysis by using a weighted GEE estimator. The weights have to be estimated using a model for $\Pr(R_i|Y_i, X_i)$, hence the non-response model must be explicitly specified and estimated, although the distribution of the error terms need not be. These weighting methods are reviewed in Section 17.5 and discussed in greater detail in Chapter 18.

Finally, when longitudinal responses are NMAR, almost all standard methods of longitudinal analysis are not valid. Both GEE methods and standard likelihood-based methods (that ignore the missing data mechanism) yield biased estimates of mean response trends. To obtain valid estimates, joint models for the response and the missing data mechanism are required. Indeed, the term *nonignorable* is used to emphasize that the missing data mechanism must be correctly specified (i.e., cannot be ignored) for inferences about the complete responses. We must also stress that any assumptions made about non-response being NMAR are completely unverifiable from the data at hand. That is, without external (or auxiliary) information about the reasons for missingness or the missingness mechanism, the observed data provide no information that can either support or refute one NMAR mechanism over another. So, short of tracking down the missing data, any assumptions made about the missingness process are not verifiable. Therefore, when missingness is thought to be NMAR, it is important to carefully assess the sensitivity of inferences to a variety of plausible assumptions concerning the missingness process. However, sensitivity analysis under different assumptions about NMAR missingness is a topic that goes well beyond the scope of this chapter.

17.4 DROPOUT

As mentioned earlier, most longitudinal studies are designed to collect data on every individual in the sample at a planned sequence of occasions. However, longitudinal studies habitually suffer from the problem of attrition; that is, some individuals "drop

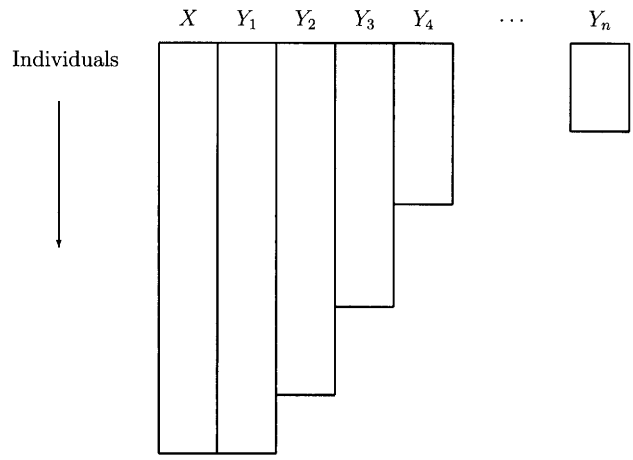

Fig. 17.1 Schematic representation of a monotone missing data pattern for dropout, with Y_j more observed than Y_{j+1} for $j = 1, ..., n-1$. Each row represents an individual; the bars represent the subset of individuals with observations on Y_j.

out" of the study prematurely. The term *dropout* refers to the special case where if Y_{ik} is missing, then $Y_{ik+1}, ..., Y_{in}$ are also missing. Alternatively, when expressed in terms of the response indicators, dropout refers to the case where if $R_{ik} = 0$ then $R_{ik+1} = \cdots = R_{in} = 0$. This gives rise to the monotone missing data pattern displayed in Figure 17.1, in contrast to the non-monotone patterns that can arise when data are missing intermittently. Note that intermittent missing data give rise to a considerably larger number of potential missing data patterns but, apart from that, do not raise any further technical considerations. As a result the focus of the remainder of this chapter is on dropout.

When there is dropout in a longitudinal study, the key issue is whether those who "drop out" and those who remain in the study differ in any further relevant way. If they do not, then analyses restricted to those remaining in the study yield valid, albeit inefficient, inferences. If they do differ, then such "complete-case" analyses are potentially biased.

In the previous section three different types of missing data mechanisms were distinguished. The same taxonomy can be applied to dropout. That is, dropout can be *completely at random*, *at random*, or *not at random*. When dropout is completely at random the probability of dropout at each occasion is independent of all past, current, and future outcomes (given the covariates). With completely random dropout, an individual leaves the study by a process unrelated to that individual's outcomes. In contrast, when dropout is at random, the probability of dropout at each occasion can depend on the previously observed outcomes up to, but not including, the current occasion. However, given the observed outcomes, dropout is assumed to

be independent of the current and future outcomes. That is, with random dropout the process can depend on the outcomes that have been observed in the past, but given this information, it is unrelated to all future (unobserved) values of the outcome variable following dropout. Finally, when dropout is not at random, the probability of dropping out at each occasion can depend on current and future unobserved outcomes. That is, dropout is said to be not at random when the process depends on the unrecorded values of the outcome variable that would have been observed had the individual remained in the study. In the context of dropout in a longitudinal study, the term "informative" dropout often is used to refer to dropout that is NMAR (similarly "non-informative" dropout often is used to refer to dropout that is either random or completely random). Here the fact of dropout is informative about the distribution of future observations. For example, consider two subjects with the same past history of responses (and covariates) up to time t. One drops out and the other does not. With MAR, their future observations have the same distribution. In contrast, dropout that is NMAR informs us that the distributions of the future observations will differ. In the general case, nothing in the data can be used to determine the distribution of the future observations of the dropouts; hence the analysis depends strongly on the specification of $\Pr(R_i|Y_i, X_i)$.

Illustration

To emphasize the main distinctions between the three types of dropout mechanism, and their potential impact on a longitudinal analysis, we consider the following simple illustration. Suppose that repeated measurements, Y_{it} $(i = 1, ..., N; t = 1, ..., 5)$, are generated from a multivariate normal distribution with mean response

$$E(Y_{it}) = \mu_{it} = \beta_1 + \beta_2 t$$

and covariance

$$\text{Cov}(Y_{is}, Y_{it}) = \rho^{|s-t|}, \quad \text{for } \rho \geq 0.$$

That is, the variance at each occasion is 1 and assumed to be constant over time, while the correlations have a first-order autoregressive pattern (see Section 7.4). Figure 17.2 displays sample means of simulated data from this model, with $N = 1000$, $\beta_1 = 5$, $\beta_2 = 0.25$, and $\rho = 0.7$. The sample means show a clear increasing trend over time and virtually coincide with the population regression line (the solid line in Figure 17.2).

Next suppose that there is dropout. When there is dropout, we can replace the vector of response indicators, R_{it} $(t = 1, ..., 5)$, with a simple dropout indicator variable, D_i, for each individual. The random variable D_i is recorded for all individuals and $D_i = k$ if an individual drops out between the $(k-1)^{th}$ and k^{th} occasion; that is, only the first $D_i - 1$ responses are observed. Assume that

$$\log\left\{\frac{\Pr(D_i = k|D_i \geq k, Y_{i1}, ..., Y_{ik})}{\Pr(D_i > k|D_i \geq k, Y_{i1}, ..., Y_{ik})}\right\} = \theta_1 + \theta_2 (Y_{ik-1} - \mu_{ik-1}) + \theta_3 (Y_{ik} - \mu_{ik}).$$

This model specifies that the probability of dropout at any occasion, given dropout has not previously occurred, can depend on the current value and the prior value of

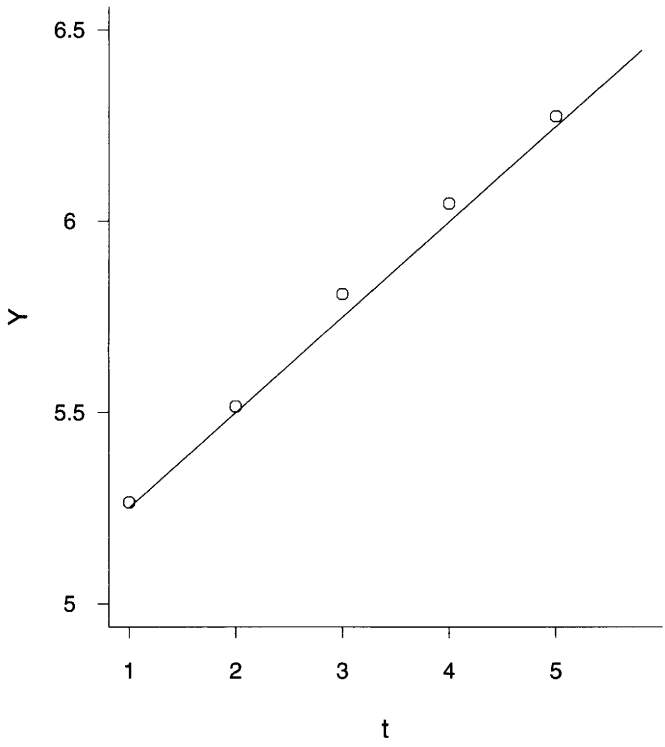

Fig. 17.2 Population regression line and empirical means at each occasion for simulated complete data.

the response variable (relative to its mean). We assume that the first response, Y_{i1}, is fully observed; that is, $\Pr(D_i = 1|D_i \geq 1, Y_{i1}) = 0$. In terms of this model for dropout, consider the following three missing data mechanisms:

(a) Dropout is MCAR: $\theta_2 = \theta_3 = 0$.

(b) Dropout is MAR: $\theta_3 = 0$. and

(c) Dropout is NMAR: $\theta_3 \neq 0$.

Figure 17.3(a) displays simulated data from this model (with $N = 1000$, $\beta_1 = 5$, $\beta_2 = 0.25$, $\rho = 0.7$) when dropout is MCAR (with $\theta_1 = -0.5$, and $\theta_2 = \theta_3 = 0$). The conditional probability of dropout at the second through fifth occasions is 0.38 (or $\frac{e^{-0.5}}{1+e^{-0.5}}$). This results in approximately 38% of the responses being missing at the second occasion, 61% missing at the third occasion, 76% missing at the fourth occasion, and 85% missing at the fifth occasion. Despite the large proportion of missing data, the empirical means at each occasion show a clear linearly increasing trend over

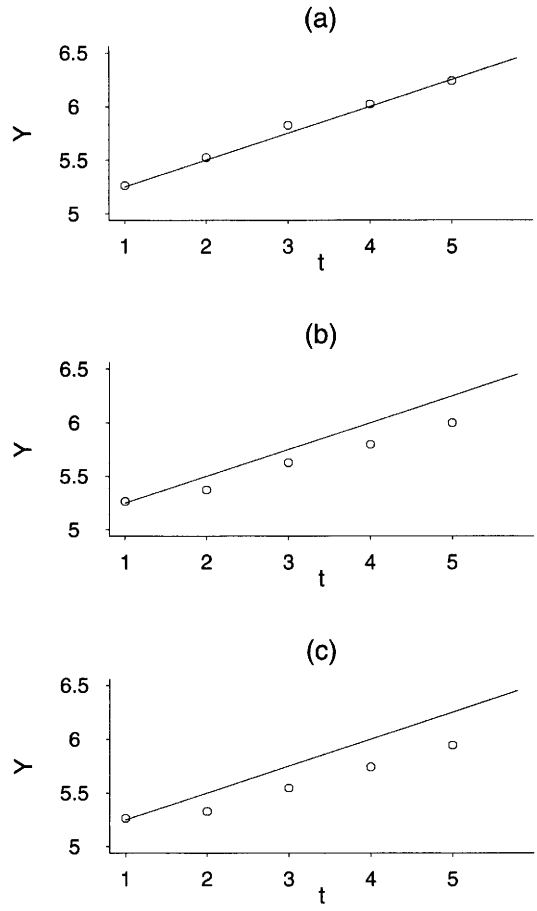

Fig. 17.3 Population regression line and observed data means at each occasion for simulated data when dropout is (a) completely at random (MCAR), (b) at random (MAR), and (c) not at random (NMAR).

time and almost coincide with the population regression line (solid line). Recall that when the missing data mechanism is MCAR, individuals with missing data are a random subset of the sample, and no bias will arise with almost any method of analysis of the observed data (either the complete data or the available data). That is, all of the methods that have been discussed so far will yield valid inferences when missing data are MCAR. To reinforce this point, the estimates of the regression parameters obtained using maximum likelihood (ML) estimation, with correctly specified covariance structure, and using a "working independence" GEE estimator are displayed at the top of Table 17.2. Recall that for a linear model with a "working independence" assumption for the covariance (and a single dispersion parameter), the GEE estimator

Table 17.2 Parameter estimates and standard errors for correctly specified likelihood analysis (ML) and "working independence" analysis (OLS/GEE) based on simulated data when dropout is (a) completely at random, (b) at random, and (c) not at random. The true regression parameters are $\beta_1 = 5.0$ and $\beta_2 = 0.25$.

		ML		OLS/GEE	
Dropout	Parameter	Estimate	SE	Estimate	SE[a]
MCAR	Intercept	5.015	0.031	5.022	0.032
	t	0.257	0.016	0.253	0.018
MAR	Intercept	5.003	0.041	5.062	0.043
	t	0.261	0.016	0.182	0.018
NMAR	Intercept	5.058	0.040	5.071	0.043
	t	0.201	0.016	0.162	0.018

[a] Standard errors for OLS/GEE are based on sandwich variance estimator.

is identical to the ordinary least squares (OLS) estimator. As expected, both the ML and OLS (or "working independence" GEE) estimates of the intercept and slope are very close to the true values of the population parameters used to generate the data. The minor differences are simply due to sampling variability.

Figure 17.3(b) displays simulated data from the same model (with $N = 1000$, $\beta_1 = 5$, $\beta_2 = 0.25$, $\rho = 0.7$) when dropout is MAR (with $\theta_1 = -0.5$, $\theta_2 = 0.5$, and $\theta_3 = 0$). Here dropout at any occasion depends on the previous response but not the current response. Because those with large values (i.e., with previous response $Y_{ik-1} > \mu_{ik-1}$) are more likely to drop out ($\theta_2 > 0$), the empirical means at the second through fifth occasions are discernibly lower than the population regression line. As a result available-data methods such as the GEE will yield biased estimates of mean response trends. In contrast, likelihood-based methods will yield valid estimates when missing data are MAR (or MCAR) and the model for the covariance has been correctly specified. The ML and GEE estimates of the intercept and slope are displayed in the middle of Table 17.2. The ML estimates are very close to the population parameters and only differ due to sampling variability. On the other hand, the "working independence" GEE (or OLS) estimate of the slope shows very discernible bias and underestimates the rate of change over time ($\widehat{\beta}_2 = 0.18$ versus $\beta_2 = 0.25$).

Finally, Figure 17.3(c) displays simulated data from the same model (with $N = 1000$, $\beta_1 = 5$, $\beta_2 = 0.25$, $\rho = 0.7$) when dropout is NMAR (with $\theta_1 = -0.5$, $\theta_2 = 0$, and $\theta_3 = 0.5$). Here dropout at any occasion depends on the current value of

the response. Because those with large values at a given occasion are more likely to be unobserved at that occasion ($\theta_3 > 0$), the empirical means are discernibly lower than the population regression line. As a result available-data methods such as the GEE yield biased estimates of mean response trends. Furthermore likelihood-based methods that ignore the missing data mechanism also yield biased estimates of mean response trends. To reinforce this point, the ML and GEE estimates of the regression parameters are displayed at the bottom of Table 17.2. The ML and GEE estimates of the slope show large biases. Of note, the magnitude of the bias is somewhat smaller for ML; however, this cannot be expected in general unless the correlation among the responses is very high. When the correlation among the responses is very high and dropout at any occasion depends only on the current value of the response, the dropout mechanism can often be approximated by an ignorable dropout mechanism that conditions on all previously observed responses

$$\Pr(D_i = k | D_i \geq k, Y_{ik}) \approx \Pr(D_i = k | D_i \geq k, Y_{i1}, ..., Y_{ik-1}).$$

For example, when the data are simulated from the same model but with a correlation parameter $\rho = 0.9$ instead of $\rho = 0.7$, the ML estimate of the slope is $\widehat{\beta}_2 = 0.24$ and the magnitude of the bias is significantly reduced. In contrast, the OLS/GEE estimate of the slope is $\widehat{\beta}_2 = 0.11$ and remains highly biased under this NMAR dropout mechanism.

17.5 COMMON APPROACHES FOR HANDLING DROPOUT

In this section we present a short review of some of the most commonly used methods for handling dropout in longitudinal analysis. We also discuss the assumptions about dropout required for each of the methods to yield valid inferences. We note that many traditional methods for handling missing data (e.g., complete-case analysis, imputation) became popular when the only approaches for analyzing data were ones based on complete and balanced data.

Complete-Case Analysis

One approach to handling dropout is to simply exclude all data from the analysis on any subject who drops out. That is, a so-called complete-case analysis can be performed by excluding any subjects that do not have data at all intended measurement occasions. We must stress that this method is very problematic and is rarely an acceptable approach to the analysis. It will yield unbiased estimates of mean response trends only when it can be assumed that dropout is MCAR. Recall that when dropout is MCAR, the study "completers" are a random subsample of the original sample from the population. However, even in cases where the MCAR assumption might be tenable, a complete-case analysis is very unappealing because of the reduction in the number of subjects contributing to the analysis. A complete-case analysis can be immensely inefficient, leading to an analysis with reduced statistical power.

Available-Data Analysis

Another approach for handling dropout is the available-data method. This is not a single method, but a very general term that refers to a wide collection of techniques that can readily incorporate vectors of repeated measures of unequal length in the analysis. For example, standard applications of generalized least squares (GLS) or the generalized estimating equations approach can be considered available-data methods, since these approaches base the analysis on all of the available observations. In general, available-data methods are more efficient than complete-case methods because they incorporate the partial information obtained from those who dropout. However, many available-data methods will yield valid analyses only if the conditional (i.e., conditional on X_i) means and covariances of the observed components of Y_i among those who dropout coincide with the corresponding conditional means and covariances of Y_i in the target population. As a result available-data methods will yield biased estimates of mean response trends unless dropout is MCAR. In general, for available-data (and complete-case) methods to be valid we require that dropout is MCAR.

Imputation

A third approach, and one that is widely used in practice, is some form of imputation for the missing responses following dropout. The idea behind imputation is very simple: substitute or fill in the values that were not recorded with imputed values. One of the chief attractions of imputation methods is that, once a filled-in data set has been constructed, standard methods for complete data can be applied. However, methods that rely on just a single imputation, creating only a single filled-in data set, fail to acknowledge the uncertainty inherent in the imputation of missing data. Multiple imputation circumvents this difficulty. In multiple imputation the missing values are replaced by a set of m plausible values, thereby acknowledging the uncertainty about what values to impute for the missing responses. The m filled-in data sets produce m different sets of parameter estimates and their standard errors. These are then appropriately combined to provide a single estimate of the parameters of interest, together with standard errors that reflect the uncertainty inherent in the imputation of the unobserved responses. Typically a small number of imputations, for instance, $10 \leq m \leq 25$, is sufficient to obtain realistic estimates of the sampling variability. A more detailed description of multiple imputation is given in Chapter 18.

Although the main idea behind imputation is very simple; what is less clear-cut is how to produce the imputed values for the missing responses. Next we consider some of the commonly used methods for imputing missing data. One widely used imputation method, especially in longitudinal clinical trials, is "last value carried forward" (LVCF), occasionally referred to as "last observation carried forward" (LOCF). This is a single imputation method that fills-in or imputes the missing values following dropout with the last observed value for that subject. Despite its widespread use, it should be recognized that LVCF makes a strong, and often very unrealistic, assumption that the responses following dropout remain constant at the last observed value

prior to dropout. Perhaps the only setting where this assumption might conceivably be appropriate is when dropout is due to recovery or cure. In the context of placebo-controlled longitudinal clinical trials, there appears to be some statistical folklore that LVCF yields a *conservative* estimate of the comparison of an active treatment versus the control. However, this is a gross misconception, and will only be true to the extent that the active treatment prior to dropout has carry-over effects following dropout. In many clinical trials this is unlikely to be the case; instead, dropout from the active treatment (e.g., due to adverse side effects) might very well result in a deterioration of the response.

Despite frequent and well-founded criticisms by statisticians, LVCF is still widely used to handle dropouts in clinical trials. Regulatory agencies such as the U.S. Food and Drug Administration (FDA) seem to encourage the continuing use of LVCF as a method for handling dropouts, despite all of its obvious shortcomings. Except in very rare cases (as mentioned above), we do not recommend the use of LVCF as a method for handling dropout. In Section 17.6 we provide an illustration of the bias that can arise when LVCF is used to impute missing values, highlighting that LVCF does not necessarily yield a *conservative* estimate of the comparison of an active treatment versus the control; for readers who find the level of detail in this section challenging, Section 17.6 can be omitted at first reading without loss of continuity.

Variations on the LVCF theme include baseline value carried forward and worst value carried forward. Worst value carried forward is most often used in comparisons of an active treatment to a placebo, since it is assumed to be conservative in that setting. However, both of these alternatives suffer the same difficulties as LVCF and cannot be counted on to give unbiased treatment estimates. In addition all of the methods suffer from optimistic standard error estimates. It is easy to see that these analyses give smaller standard errors than complete-case, or even available-data estimates because they assume complete data on everyone. However, they will generally give smaller standard errors than what we would expect if we had been fortunate enough to have complete data on everyone. This is because the variability of baseline measurements is usually smaller because of selection criteria into the study, and as we move out in time, the observations tend to become more variable. Hence substituting baseline or intermediate values for final values can be expected to give a less variable data set. It is also true if we use worst value, since worst values are often similar especially for responses based on a scale. Therefore we caution that neither LVCF nor any of its variants, such as baseline value carried forward, provide a legitimate approach for analyzing incomplete data. These ad hoc methods of imputation typically produce bias whose direction and magnitude depend on both the true, but unknown, treatment effect and the dropout rates in the treatment groups (see Section 17.6). In addition, similar to other single imputation methods, these methods artificially increase the amount of information in the data by regarding imputed and actually observed values on an equal footing.

Other imputation methods that have a much firmer theoretical foundation draw values of Y_i^M from the conditional distribution of the missing responses given the observed responses, $f(Y_i^M|Y_i^O, X_i)$. With the monotone missing data patterns produced by dropouts, it is relatively straightforward to impute missing values by drawing

values of Y_i^M from $f(Y_i^M|Y_i^O, X_i)$ in a sequential manner. A variety of imputation methods can be used to draw values from $f(Y_i^M|Y_i^O, X_i)$; we describe two distinct methods in our discussion of multiple imputation in Chapter 18. When missing values are imputed from $f(Y_i^M|Y_i^O, X_i)$, regardless of the particular imputation method adopted, subsequent analyses of the observed and imputed data are valid for dropouts that are MAR (or MCAR). Furthermore multiple imputation ensures that the uncertainty is properly accounted for.

Finally, there is another related form of imputation where the missing responses are effectively imputed by modeling and estimating parameters for the joint distribution of Y_i, $f(Y_i|X_i)$. When dropout is MCAR or MAR, likelihood-based methods can be used based solely on the marginal distribution of the observed data. If dropout is MCAR or MAR (and the parameters of the dropout and outcome processes are distinct, a technical requirement that can usually be assumed in practice), then ML estimates can be obtained by maximizing $f(Y_i^O|X_i)$, where $f(Y_i^O|X_i)$ denotes the ordinary marginal distribution of the particular subset of Y_i determined by Y_i^O. Importantly, likelihood-based inference does not require specification of the dropout mechanism and the contribution of $\Pr(D_i|Y_i^O, X_i)$ to the likelihood can be ignored. In a certain sense, the missing values are validly predicted by the observed data via the model for the conditional mean, $E(Y_i^M|Y_i^O, X_i)$. In many missing data situations, ML estimates of the parameters can be easily obtained by an iterative *EM algorithm* that alternates between filling in missing values (the expectation or E-step), then maximizing the likelihood for the resulting filled in data set (the maximization or M-step). For example, if the responses are assumed to have a multivariate normal distribution, then predictions of the missing values are based on the conditional mean of Y_i^M, given Y_i^O (and X_i),

$$E(Y_i^M|Y_i^O, X_i) = \mu_i^M + \Sigma_i^{MO}{\Sigma_i^O}^{-1}(Y_i^O - \mu_i^O),$$

where μ_i^M and μ_i^O denote those components of the mean response vector corresponding to Y_i^M and Y_i^O, and Σ^O and Σ_i^{MO} denote those components of the covariance matrix corresponding to the covariance among the elements of Y_i^O and the covariance between Y_i^M and Y_i^O. This simple implementation of the EM algorithm works for estimating means in the setting of the multivariate normal distribution, but for estimating variances or covariances, somewhat more complex expressions are required for filling in the missing observations.

Weighting Methods

An alternative approach for handling dropout is to weight the observed data in some appropriate way. In weighting methods, the under-representation of certain response profiles in the observed data is taken into account and corrected. A variety of different weighting methods that adjust for dropout have been proposed. These approaches are often called propensity weighted or inverse probability weighted (IPW) methods. Here the underlying idea is to base estimation on the observed responses but weight them to account for the probability of remaining in the study. The probability of

remaining in the study can be estimated as a function of the observed responses prior to dropout, and also as a function of the covariates and any additional variables or subject characteristics that are thought likely to predict dropout.

In the simplest version of this approach a single weight, w_i, is calculated only for those individuals who complete the study. The weight for each individual denotes the inverse probability of remaining in the study until the last intended measurement occasion; that is, $w_i = \{\Pr(D_i = n+1)\}^{-1}$. It can be computed sequentially as the inverse of the following product of the conditional probabilities of remaining in the study at each occasion:

$$\begin{aligned} w_i &= \{\Pr(D_i = n+1)\}^{-1} \\ &= \{\Pr(D_i > 1 | D_i \geq 1) \times \Pr(D_i > 2 | D_i \geq 2) \times \cdots \times \Pr(D_i > n | D_i \geq n)\}^{-1} \\ &= (\pi_{i1} \times \pi_{i2} \times \cdots \times \pi_{in})^{-1}, \end{aligned}$$

where $\pi_{ik} = \Pr(D_i > k | D_i \geq k)$ can be estimated from those remaining at the $(k-1)^{st}$ occasion, given the recorded history of all available data up to the $(k-1)^{st}$ occasion. Given the estimated weight, $\widehat{w}_i$, a weighted complete-case analysis can be performed. For example, the GEE approach can be adapted to handle data that are MAR by making adjustments to the analysis for the probability of remaining in the study. One variant of this approach is to use a weighted GEE to analyze the data from the study "completers", with weights inversely proportional to the estimated probability that the i^{th} subject completes the study. In the weighted complete-case analysis each subject's contribution to the analysis is weighted by $\widehat{w}_i$, thereby providing valid estimates of the mean response trends when dropout is MAR.

Inverse probability weighted methods were first proposed in the sample survey literature where the weights are known and based on the survey design. In the sample survey setting, units are sampled with unequal probability of selection and therefore must be given correspondingly unequal weights (inverse probability weighting) in the analysis. In the missing data setting, the intuition behind the weighting methods is that each subject's contribution to the weighted complete-case analysis is replicated w_i times, in order to count once for herself and $(w_i - 1)$ times for those subjects with the same history of responses and covariates who do not complete the study. These weights correct for the under-representation of certain response profiles in the observed data due to dropout. The weighting methods are valid provided that the model that produces the estimated w_i is correctly specified.

In longitudinal analyses, w_i is not ordinarily known, but must be estimated from the observed data (e.g., using a repeated sequence of logistic regressions for the π_{ik}'s). Therefore the variance of inverse probability weighted estimators should also account for estimation of w_i. Counter-intuitively, estimation of the weights from the data at hand leads to improvements in precision. Finally, we note that this approach for handling dropout can be made more efficient by conducting an appropriately weighted available-data analysis. This requires that occasion-specific weights for each individual, w_{ij}, be incorporated into the analysis, where w_{ij} denotes the inverse probability that the i^{th} subject is still in the study at the j^{th} occasion. A more detailed description of inverse probability weighted methods is given in Chapter 18.

17.6 BIAS OF LAST VALUE CARRIED FORWARD IMPUTATION*

In Section 17.5 we noted that last value carried forward (LVCF) imputation makes a strong, and often very unrealistic, assumption that the responses following dropout remain constant at the last observed value prior to dropout. In this section[1] we provide a demonstration of the bias that can arise when LVCF is used to impute missing values, highlighting that LVCF does not necessarily yield a *conservative* estimate of the comparison of an active treatment to the control. When LVCF is used in the comparison of an active treatment to a control, the bias can go in either direction. The content of this section is somewhat technical and can be omitted on first reading of this chapter without loss of continuity.

The three mechanisms, MCAR, MAR, and NMAR, introduced in Section 17.2 are only one approach to modeling missing data, albeit the most widely used approach. There is another approach known as *pattern mixture models* that can be easier to understand and specify in certain missing data situations, especially when modeling dropout. Simply put, pattern mixture models stratify the data by missing data patterns and assume a different model for the data within each stratum (e.g., assume that subjects who drop out are those who are likely not responding well to treatment). To study the potential bias of LVCF imputation, it is easier to use a pattern mixture model approach to determine what kind of bias might accrue.

To illustrate the potential bias that can arise from LVCF, consider a simple two-group design (e.g., active treatment versus control) with two repeated measures of the response, one at baseline, the other at end of follow-up. Let $\text{trt}_i = 1$ if the i^{th} subject is assigned to the active treatment group and $\text{trt}_i = 0$ if assigned to the control group. We assume that the baseline response, Y_{i1}, is always observed ($R_{i1} = 1$ for all individuals) and the probability of the follow-up response, Y_{i2}, being observed is $\pi_0 = \Pr(R_{i2} = 1|\text{trt}_i = 0)$ for those in the control group and $\pi_1 = \Pr(R_{i2} = 1|\text{trt}_i = 1)$ for those in the active treatment group. When stratified in terms of being a "dropout" (with $R_{i2} = 0$) or a "completer" (with $R_{i2} = 1$), two saturated (or unrestricted) models for the change in the mean response can be specified as

$$E(Y_{i2} - Y_{i1}|\text{trt}_i, R_{i2} = 0) = \alpha_1 + \alpha_2 \text{trt}_i, \tag{17.4}$$

$$E(Y_{i2} - Y_{i1}|\text{trt}_i, R_{i2} = 1) = \gamma_1 + \gamma_2 \text{trt}_i. \tag{17.5}$$

The alert reader would have recognized that without some assumptions about the missing data mechanism, α_1 and α_2 in (17.4) cannot be estimated from the available data because Y_{i2} is not observed for those who dropout. Recall from Section 17.2 that for the special case where the missing data mechanism is assumed to be MCAR, the distribution of Y_i (given the covariates) is the same in each of the distinct subpopulations defined by the missing data patterns. In this illustration there are only two

[1]This section derives formulae for the potential bias of LVCF in a simple setting and is based on similar derivations in Chapter 27 of Molenberghs and Verbeke (2005).

sub-populations, the "dropouts" and the "completers." Therefore, under the MCAR assumption, $\gamma_k = \alpha_k$ for $k = 1, 2$ in equations (17.4) and (17.5).

In such a study design, the primary goal of the analysis is to compare the two groups in terms of their changes in response from baseline to follow-up. Specifically, the parameter of primary interest can be expressed as

$$\delta = E(Y_{i2} - Y_{i1}|\text{trt}_i = 1) - E(Y_{i2} - Y_{i1}|\text{trt}_i = 0). \tag{17.6}$$

Note that the parameter δ is expressed in terms of $E(Y_{i2} - Y_{i1}|\text{trt}_i)$ not $E(Y_{i2} - Y_{i1}|\text{trt}_i, R_{i2})$. We can obtain $E(Y_{i2} - Y_{i1}|\text{trt}_i)$ by simply averaging $E(Y_{i2} - Y_{i1}|\text{trt}_i, R_{i2})$ over the distribution of R_{i2} (given treatment group). Specifically,

$$\begin{aligned} E(Y_{i2} - Y_{i1}|\text{trt}_i) &= E(Y_{i2} - Y_{i1}|\text{trt}_i, R_{i2} = 0)\Pr(R_{i2} = 0|\text{trt}_i) \\ &\quad + E(Y_{i2} - Y_{i1}|\text{trt}_i, R_{i2} = 1)\Pr(R_{i2} = 1|\text{trt}_i). \end{aligned}$$

From equations (17.4) and (17.5), it can be seen that

$$\begin{aligned} E(Y_{i2} - Y_{i1}|\text{trt}_i = 1) &= E(Y_{i2} - Y_{i1}|\text{trt}_i = 1, R_{i2} = 0)\Pr(R_{i2} = 0|\text{trt}_i = 1) \\ &\quad + E(Y_{i2} - Y_{i1}|\text{trt}_i = 1, R_{i2} = 1)\Pr(R_{i2} = 1|\text{trt}_i = 1) \\ &= E(Y_{i2} - Y_{i1}|\text{trt}_i = 1, R_{i2} = 0)(1 - \pi_1) \\ &\quad + E(Y_{i2} - Y_{i1}|\text{trt}_i = 1, R_{i2} = 1)\pi_1 \\ &= (\alpha_1 + \alpha_2)(1 - \pi_1) + (\gamma_1 + \gamma_2)\pi_1. \end{aligned}$$

Similarly

$$\begin{aligned} E(Y_{i2} - Y_{i1}|\text{trt}_i = 0) &= E(Y_{i2} - Y_{i1}|\text{trt}_i = 0, R_{i2} = 0)\Pr(R_{i2} = 0|\text{trt}_i = 0) \\ &\quad + E(Y_{i2} - Y_{i1}|\text{trt}_i = 0, R_{i2} = 1)\Pr(R_{i2} = 1|\text{trt}_i = 0) \\ &= E(Y_{i2} - Y_{i1}|\text{trt}_i = 0, R_{i2} = 0)(1 - \pi_0) \\ &\quad + E(Y_{i2} - Y_{i1}|\text{trt}_i = 0, R_{i2} = 1)\pi_0 \\ &= \alpha_1(1 - \pi_0) + \gamma_1\pi_0. \end{aligned}$$

Therefore

$$\begin{aligned} \delta &= \{(\alpha_1 + \alpha_2)(1 - \pi_1) + (\gamma_1 + \gamma_2)\pi_1\} \\ &\quad - \{\alpha_1(1 - \pi_0) + \gamma_1\pi_0\} \\ &= \alpha_2 + (\pi_1 - \pi_0)(\gamma_1 - \alpha_1) + \pi_1(\gamma_2 - \alpha_2). \end{aligned}$$

Next we consider the target parameter of the analysis based on LVCF imputation. Recall that in the LVCF analysis it is assumed that $E(Y_{i2}|\text{trt}_i, R_{i2} = 0) = E(Y_{i1}|\text{trt}_i, R_{i2} = 0)$; that is, $E(Y_{i2} - Y_{i1}|\text{trt}_i, R_{i2} = 0) = 0$. Under this assumption it can be shown that the target parameter for the LVCF analysis, denoted δ_{LVCF}, is

$$\delta_{LVCF} = (\gamma_1 + \gamma_2)\pi_1 - \gamma_1\pi_0 = (\pi_1 - \pi_0)\gamma_1 + \pi_1\gamma_2.$$

Thus it is transparent that $\delta_{LVCF} \neq \delta$ and that, in general, the LVCF analysis will yield a biased estimate of δ, the parameter of primary interest.

To see what kind of bias might accrue, consider the case where the underlying missing data mechanism is assumed to be MCAR. Under the MCAR assumption, $\gamma_k = \alpha_k$ for $k = 1, 2$ in equations (17.4) and (17.5), and the expression for δ simplifies to $\delta = \gamma_2 = \alpha_2$. However, even under the MCAR assumption, the target parameter of the LVCF analysis is

$$\delta_{LVCF} = (\pi_1 - \pi_0)\gamma_1 + \pi_1\gamma_2 \neq \delta.$$

Thus the LVCF analysis yields a biased estimate of δ even under MCAR. Under MCAR the amount and direction of bias for the LVCF analysis is given by

$$(\delta_{LVCF} - \delta) = (\pi_1 - \pi_0)\gamma_1 - (1 - \pi_1)\gamma_2.$$

To demonstrate that the bias can be in any direction, both positive and negative, consider the case where $\gamma_1 \neq 0$ and $\gamma_2 = 0$. This corresponds to the scenario where there is change in the mean response from baseline in both treatment groups but at the same rate (i.e., there is no differential treatment effect on the pattern of change). Then the bias can go in either direction depending on the signs of $(\pi_1 - \pi_0)$ and γ_1. This highlights how the LVCF analysis can potentially produce an apparent treatment effect, favoring either the active treatment group or the control group, when no such effect exists ($\delta = \gamma_2 = 0$). Under the assumption that the missing data are MAR, it is also possible to derive expressions for the bias of the LVCF analysis. Under MAR, expressions for the bias are somewhat more complicated but nonetheless reveal that the LVCF analysis also yields a biased treatment comparison, with bias that can operate in either direction.

In summary, although LVCF is a widely used imputation method, especially in longitudinal clinical trials, it makes a strong, and often very unrealistic, assumption about the responses following dropout. As we have seen, even when missingness can be assumed to be MCAR, LVCF yields a biased treatment comparison and the bias can go in either direction. Moreover, due to this bias, an LVCF analysis can potentially yield an apparent treatment effect when no such effect exists. In contrast, under MCAR, almost any other method of analysis, including a complete-case analysis, is unbiased. Therefore, except in very rare cases (e.g., when dropout is due to cure or recovery) we do not recommend the use of LVCF as a method for handling dropout. Finally, we note that LVCF happens to be equivalent to "baseline value carried forward" in the simple illustration used in this section. Therefore, all our criticism of LVCF applies equally to "baseline value carried forward" imputation.

17.7 FURTHER READING

A useful discussion of methods for handling dropout in longitudinal studies can be found in Heyting et al. (1992). The tutorial article by Hogan et al. (2004) provides a comprehensive overview of more recent developments in methods for adjusting for drop-out within likelihood-based and semiparametric modeling frameworks and illustrates their application with two worked examples. White et al. (2011) suggest a

general framework for "intention to treat" analysis in randomized clinical trials that depends on making plausible assumptions about the missing data and including all participants in sensitivity analyses.

In longitudinal clinical trials, "last value carried forward" (LVCF) imputations are still widely used to handle dropouts; see Ware (2003), Cook et al. (2004), Molenberghs et al. (2004), and Kenward and Molenberghs (2009) for critiques of this method and its variants (e.g., baseline value carried forward). The illustration of the potential bias that can arise from LVCF in Section 17.6 is based on similar derivations in Chapter 27 of Molenberghs and Verbeke (2005).

Bibliographic Notes

Rubin (1976) developed the taxonomy for describing the assumptions concerning the dependence of the missingness process on observed and unobserved responses. Little and Rubin (2001) is the definitive textbook on missing data, providing a comprehensive description of the theory and application of methods for handling missing data; also see Schafer (1997), Tsiatis (2006), Molenberghs and Kenward (2007), and Daniels and Hogan (2008). Laird (1988) discusses missing data issues in longitudinal studies; also see the review articles by Little (1995) and Kenward and Molenberghs (1999). The EM algorithm, a general technique for ML estimation with incomplete data, is discussed in the seminal paper by Dempster et al. (1977). Finally, Hogan and Laird (1996) and Little and Yau (1996) discuss methods for handling missing data for "intention to treat" analysis in randomized clinical trials.